

# Unsupervised Visual and Textual Information Fusion in Multimedia Retrieval - A Graph-based Point of View\*

Gabriela Csurka<sup>1</sup>, Julien Ah-Pine<sup>2</sup>, and Stéphane Clinchant<sup>1</sup>

<sup>1</sup>*Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240, Meylan France, [Firstname.Lastname@xrce.xerox.com](mailto:Firstname.Lastname@xrce.xerox.com)*

<sup>2</sup>*University of Lyon 2, ERIC Lab, 5, avenue Pierre Mendès France, 69676 Bron Cedex, France, [julien.ah-pine@eric.univ-lyon2.fr](mailto:julien.ah-pine@eric.univ-lyon2.fr)*

## Abstract

Multimedia collections are more than ever growing in size and diversity. Effective multimedia retrieval systems are thus critical to access these datasets from the end-user perspective and in a scalable way. We are interested in repositories of image/text multimedia objects and we study multimodal information fusion techniques in the context of content based multimedia information retrieval. We focus on graph based methods which have proven to provide state-of-the-art performances. We particularly examine two of such methods : cross-media similarities and random walk based scores. From a theoretical viewpoint, we propose a unifying graph based framework which encompasses the two aforementioned approaches. Our proposal allows us to highlight the core features one should consider when using a graph based technique for the combination of visual and textual information. We compare cross-media and random walk based results using three different real-world datasets. From a practical standpoint, our extended empirical analysis allow us to provide insights and guidelines about the use of graph based methods for multimodal information fusion in content based multimedia information retrieval.

## 1 Introduction

With the continuous growth of communication technologies, the information that we consult, produce and communicate whatever the communication device we use, has been richer and richer in terms of the media it is composed of. The web has particularly contributed to the production of such multimedia or multimodal data. For instance, web pages from news agencies websites are texts illustrated with pictures or videos; photo sharing websites, such as FlickrR, store pictures annotated with tags; video hosting websites, such as Youtube, are again examples of multimedia data repositories. Apart from the web, we have also witnessed the development of new services that rely on digital libraries made of data composed of several media. In museums for example, there are more and more multimedia applications using text, image, video and speech in order to better plunge the visitor into the historical context of the piece of art she is consulting. New generations of television devices now propose on-line interactive media, on-demand streaming media and so on. The ever-growing production of multimodal data has brought the multimedia research community to address the problem of effectively accessing multimedia repositories from the end-user perspective and in a scalable way. Accordingly, multimedia data search has been a very active research domain for the last decades.

There are different ways to search a multimedia repository. As for videos or images datasets such as Youtube or FlickrR, we typically index those media by means of the title, metadata, tags or text associated to or surrounding them. Then, we search those multimodal objects by using text queries and text based search engines. There are different reasons we use text to retrieve videos or images. Firstly, it is not always possible for the user to query a collection by examples, since the search engine cannot always provide her with videos or images that represent the

---

\*An extended version of the paper Visual and Textual Information Fusion in Multimedia Retrieval using Semantic Filtering and Graph based Methods, by J. Ah-Pine, G. Csurka and S. Clinchant, submitted to ACM Transactions on Information Systems.

type of items she would like to retrieve. Secondly, videos or images are stored in machines into a computational representation consisting of low-level features which do not carry by their own the high level semantics. In other words, it is a strong challenge to effectively associate low-level features extracted from videos or images with high-level features such as keywords or tags without using pre-trained classifiers. This problem is known as the semantic gap. As a consequence of those two difficulties, we generally use the text media for content based multimedia information retrieval (CBMIR) in order to have more relevant search results.

If the text is the core media to use in order to access a multimedia repository effectively, it is however beneficial to use other media in addition to the text, during the course of the search process. Indeed, most of research works about multimedia information fusion have shown that combining different modalities to address CBMIR tasks, even with simple strategies, is beneficial. In this paper we are interested in this topic. We particularly address the combination of visual and textual information. We thus deal with repositories that are composed of multimedia objects made of an image associated with a text. There are different multimedia information fusion methods and in this paper we are interested in graph based techniques. Such approaches became very popular in the information retrieval community since the development of techniques like PageRank or Hits [10, 34, 36].

In a nutshell, the goals of this paper are the following ones :

- We discuss the semantic filtering method that seeks to enhance the similarities between multimedia items when they are composed of both a visual and a textual part [16]. We explain how such a filter based on the text query can better cope with the semantic gap in the case of CBMIR. We propose to use this approach as a first level of the fusion process of visual and textual information in our multimedia relevance model. Indeed, not only the proposed semantic filtering improves the similarity measures between multimedia items but it also allows reducing the storage and computational complexities of graph based models.
- We study and compare two popular graph based multimedia information fusion methods that were originally proposed in two different research communities. On the one hand, we analyze the cross-media similarity approach initially proposed in [15, 14] for content based image retrieval in the context of Image-CLEF multimedia retrieval tasks. On the other hand, we investigate the random walk approach which was initially proposed in [31, 30] and used on several TRECVID tasks for content based video retrieval. Our main contribution in that perspective is to show that the two techniques are related. In fact, we propose a unifying framework that generalizes both approaches. This generalization allows us to better compare the two techniques, to propose a third approach which amounts to a mix of both latter methods, and it also aims at examining the main points and settings when using graph based methods for the combination of visual and textual information in CBMIR.
- We analyze two different multimodal search scenarios. In the first scenario, we suppose that the user can only use a text query in order to retrieve images. Multimedia objects of the repository are indexed using their text part and a text based search engine is used in a first time. In a second time, we use the visual information of the (text based) retrieved objects in order to improve the search results. This multimedia search scenario is referred along this paper as **the asymmetric case** since the user can only use a text query. In contrast, in the second scenario, the user can use a multimedia query which means that she can enter a text query accompanied with one or several images as examples of her information need. To this second scenario we refer as **the symmetric search scenario**.
- We experiment with 3 different image/text datasets which have distinct features. We conduct many tests in order to have a better analysis of the core points in the use of the graph based methods under study and in the context of image/text multimedia retrieval. Our experimental results allow us to provide insights and guidelines about how to set the parameters of the unified graph based technique we propose.

The rest of this paper is organized as follows. In section 2 we review the main families of multimodal fusion approaches and their features. We take into consideration both the asymmetric and the symmetric search scenarios. In section 3, we discuss the use of graph based methods to fuse visual and textual information and we detail the cross-media similarities and the random walk based techniques which are the techniques under study in this paper. Next, in section 4, we introduce the semantic filtering method which represents a core step to refine multimedia similarities from a semantic standpoint, when textual information is available. Such an approach amounts to a first level of multimedia information fusion and in addition, it also enables reducing the storage and computational complexities of graph based methods. In section 5, in light of the material exposed in

sections 3 and 4, we introduce our multimedia relevance model that relies both on the semantic filtering guided by the text query and an unifying graph based framework that embodies both cross-media similarities and random walks based scores. Then, we describe in section 6, the experimental settings we conducted on three real-world multimedia collections in order to validate our work. We introduce the image and text representations and the monomedia similarity matrices we used. In section 7, we present the experimental results we obtained with the different tested fusion strategies in the goal of comparing cross-media similarities and random walk based scores. We finally discuss some other advantages of our proposal in terms of complexity as regard to large collections and we provide some guidelines on how to use the generalized graph based approach we propose. In section 8, we summarize our main findings.

## 2 Families of unsupervised multimedia fusion techniques in CBMIR

A good introduction to the multimedia information access domain, its challenges and its basic techniques can be found in [54] that covers the common topics in multimedia IR such as feature extraction, distance measures, supervised classification also known as automatic tagging and fusion of different experts. In this paper, we are particularly interested in multimedia fusion techniques and the literature on this topic is very vast. In this section we attempt to depict the main families of fusion methods for visual and textual information. It is important to precise that we place ourselves in an unsupervised context meaning that we do not use any learning technique in our framework. We can mention at least two research communities that have been addressing this research topic actively. On the one hand, there are the research teams that have participated in the TRECVID workshop series and have focused their research efforts on video retrieval [57]. On the other hand, we can quote the research groups involved in the ImageCLEF meetings and which have been interested in the tracks related to image and multimedia retrieval [45]. In the former research community it is usually assumed that the user does not have any example query and the common way to search a multimedia collection rely solely on textual queries. On the contrary, in the latter research community, it is generally assumed that the user information need is expressed by a multimedia query composed of an image query and a related text query. We present in the following, broad families of multimedia fusion techniques that have been studied for the two distinct search scenarios.

Despite the fact that we focus on unsupervised multimedia fusion, we also point to some research papers that address multimedia fusion techniques from a supervised or a semi-supervised perspective and which show some connections with our work.

### 2.1 The symmetric case with an image query and a text query

Most of the techniques developed in this context fall in three different categories : early, late and transmedia fusion. We depict these three families of approaches by distinguishing the inherent steps they are composed of. This is summarized in Figure 1. In the following, we assume that the multimedia query can be considered similar as any item of the multimedia collection that is to say an object made of an image part and a text part. Given a multimedia query, the search process consists in measuring a multimedia similarity between the query and the multimedia items in the repository.

The early fusion approach represents the multimedia objects in a multimodal feature space designed *via* a joint model that attempts to map image based features to text based features and *vice versa*. The simplest early fusion method consists in concatenating both image and text feature representations (see *e.g.* [58, 18, 45]). However, more elaborated joint models such as Canonical Correlation Analysis have been investigated [43, 37, 60, 52]. In the same vein, [41] presents an information theoretic framework that could also fit into this family of fusion approaches.

On the contrary, late fusion and transmedia fusion strategies do not act at the level of the monomedia feature representations but rather at the level of the monomedia similarities [15, 11]. In these contexts, we assume that we have effective monomedia similarities and that it is better to combine their respective decisions rather than attempting to bridge the semantic gap at the level of the features.

Concerning late fusion techniques, they mainly seek to merge the monomedia relevance scores by means of aggregation functions. In that case, the simplest aggregation technique used is the mean average [25] but more elaborated approaches have been studied (*e.g.* [12, 45, 23, 65]).

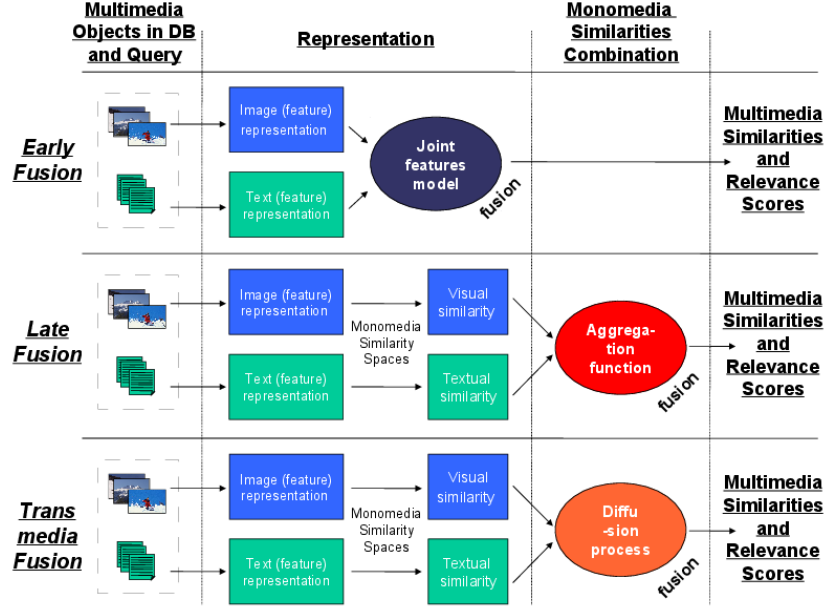


Figure 1: Early, late and transmedia fusion.

As far as transmedia fusion methods are concerned, they act like similarity diffusion processes. The resulting combination is non linear unlike most of late fusion techniques. These methods generally amount to mixing monomedia similarity matrices by means of matrix multiplication operations [64, 47, 31, 15]. In that kind of relevance models, we usually start the diffusion process using pseudo-relevant items only. In that case, we typically use the  $k$  nearest neighbors according to one monomedia relevance score and thus those methods are also inspired from the pseudo-relevance feedback mechanism in information retrieval (see *e.g.* [55]).

It is important to mention that there are other ways to categorize the different multimedia fusion techniques. In the recent survey paper [67] for example, other terms are used. Nevertheless, they basically correspond to the definitions given above with the following mappings : early, late and transmedia fusion are named latent space based, linear fusion and graph based fusion in [67].

## 2.2 The asymmetric case with a text query only

In addition to the three previously recalled types of fusion methods, [67] cites another category named visual reranking. This fourth family of techniques assumes that multimedia collections are accessed using textual queries solely. Therefore, in this context, there is an explicit asymmetry between image and text in the multimedia search scenario.

Visual reranking techniques particularly deals with such a search scenario. They proceed in two steps : using the text query, they first use text based similarities in order to find the most relevant objects from a semantic viewpoint; then, they employ the visual similarities between objects of the database in order to refine the textual similarities based ranking. Similarly to Figure 1, we depict in Figure 2 the different main steps of visual reranking approaches.

The common assumption that all visual reranking techniques make is that visually similar images should have similar relevance scores [44]. However, different approaches are used to re-arrange the top retrieved items by the text similarities in order to take this principle into account. According to [67], we can categorize visual reranking techniques into three subcategories : classification based, clustering based and graph based.

In the first case, pseudo-positive and pseudo-negative objects are sampled from the text based ranked list then a learning to rank algorithm is trained on the visual features (see *e.g.* [39] for a general reference on learning to rank methods). Afterward, objects are re-ordered according to the scores provided by the trained classifier. The critical point is the sampling method used to select pseudo-training examples. The simplest strategy considers

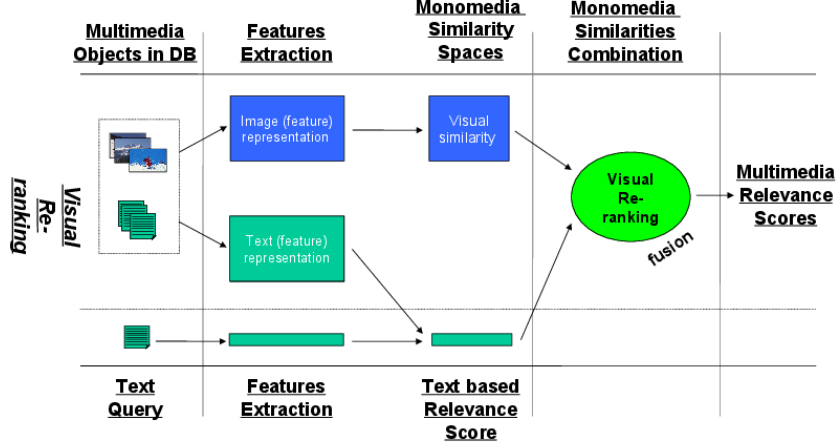


Figure 2: Visual reranking.

items at the top of the list as pseudo-positive and items at the bottom as pseudo-negative but more sophisticated approaches have been proposed [59, 66, 44].

As for clustering based visual reranking, the main idea is to cluster the list of text based retrieved items and to re-arrange them such that objects that are visually highly similar and have high initial text retrieval scores are favored [29, 30].

Graph based methods consider multimedia objects as nodes of a graph and the different types of relationships they share as edges. Examples of weighted edges between objects are visual similarities or textual similarities but depending on the application other types of relations can be considered. Graph analysis techniques are then employed in order to infer new features in the goal of re-arranging the text based ranked list of items. One such method, inspired by the well-known PageRank [9, 36, 26] used to rerank web pages by search engines such as Google, was proposed in [31, 30]. It is based on random walks over a stochastic matrix which is deduced from the fusion of visual and textual similarities, and the stationary probability distribution over the nodes is then additionally used to rerank the initial retrieved list. In the same vein, [21] proposed a Markov random walk model with backward and forward steps. They found out that the best performances were obtained with a long backward walk with high self-transition probability.

### 2.3 Graph based techniques in both search scenarios

Transmedia fusion techniques we introduced in paragraph 2.1 are technically similar to graph based methods presented in the previous paragraph. Indeed, both approaches use similarity matrices to respectively rank or rerank multimedia items. Graph based methods have proven to be state-of-the-art techniques for many information retrieval tasks (see *e.g.* [9, 36, 26]). In CBMIR too, they have demonstrated their advantages over early or late fusion approaches in many research works (see *e.g.* [45, 67]). We thus focus on such methods in this paper.

Besides, there has been very few research works that address CBMIR in a symmetric search scenario and using graph based methods. Consequently, in this paper we study the different kinds of search scenarios with such techniques in order to have a better comparison between the asymmetric and the symmetric search scenarios in this context.

Before presenting in more details the two graph based fusion techniques we examine in the rest of the paper, we present in the next paragraph some additional references that also tackle multimedia information fusion and/or multimedia retrieval but in other learning settings.

### 2.4 Multimedia fusion in a supervised or a semi-supervised context

We review some related research papers that tackle video and image search from a multimodal perspective but employing supervised or semi-supervised techniques. In [27] for example, the authors use hypergraph learning

to design a joint visual-textual representation of multimedia objects. This method amounts to an early fusion scheme. Another early fusion approach was presented in [46] and which addresses multimedia query expansion for both the text and the image parts. This work relies on an intermediate representation of multimedia information in a predefined visual-concept lexicon. Classification models are used to map the queries to the lexicon. Then based on pseudo-relevance feedback different query expansion and score reranking methods are proposed. Similarly, [53] uses intermediate representation, in their case visual classifiers. To build these classifiers they download images from Google or Bing using query words and represent these images by classes (attribute-based image descriptors). In a second step, images in the web pages are classified using these classifiers and the scores are used to rerank the multimodal documents (in their cases the web pages). The reranking is also supervised as a set of training queries with relevance scores are used to learn the parameters of the latter algorithm.

Other related works that are worth mentioning are the following ones [62, 61, 63]. These papers address video semantic annotation and web image search in a semi-supervised fashion. The general framework used in these contributions is formulated as an optimization problem that simultaneously deal with the late fusion of monomedia similarity matrices and graph semi-supervised learning. The solutions of the optimization problems can be formulated using normalized graphs Laplacian and iterated algorithms are proposed to infer the relevance scores which are further used for annotating videos or ranking images.

The main differences between these research works and our framework are the following ones : (i) we do not use any learning models nor external resources (such as a domain ontology or downloaded image set) and we only rely on the surrounding text of images which is a more general setting; (ii) we emphasize the transmedia principle in the diffusion process which mix the monomedia similarity matrices and relevance scores differently from late fusion; (iii) since no learning phase is required in our case we avoid the annotation burden and also the time complexity problem underlying such methods.

After having introduced a classification of the most used unsupervised multimedia information fusion strategies and discussed some other related works, we introduce in the next section, the graph based fusion methods we are going to embed in our multimedia relevance model.

### 3 Cross-media similarities and random walk based scores

We recall two popular image/text graph based fusion techniques in CBMIR and we consider their use in the two different search scenarios we recall previously. The first approach called cross-media similarities was proposed in the context of ImageCLEF workshop series while the second method based on random walks and called context reranking was used in TRECVID tasks.

For convenience, we introduce in Table 1 the notations we will use in the rest of the paper. Note that we assume that the different similarities or scores are all non negative numbers.

#### 3.1 Methods based on cross-media similarities

Cross-media similarities studied in this paper refer to the research work developed in the following references [15, 1] and which has proven to give top-ranked retrieval results on several ImageCLEF multimedia search tasks<sup>1</sup> [45].

We can explain the cross-media similarity mechanism using the following illustration (see also Figure 3). Given a text query  $q_t$ , we first find the most similar items in the collection with regard to the textual similarities. Then, we select pseudo-relevant objects  $d$  which are the set of  $k$  nearest neighbors. Next, we look at the pseudo-relevant objects' visual similarities profiles  $S_v(d, \cdot)$ . We then combine these visual similarity scores linearly and we obtain a cross-media similarity measure between the text query and the multimedia objects of the database. Formally such cross-media similarities are defined as follows :

$$cm_{tv}(q, \cdot) = \mathbf{K}(s_t(q, \cdot), k) \cdot S_v \quad (1)$$

where :

- $\mathbf{K}(\cdot, k)$  is an operator that takes as input a vector and gives a zero value to elements whose score is strictly lower then the  $k^{th}$  highest score.

<sup>1</sup>For more details, please visit [www.imageclef.org](http://www.imageclef.org)

Notations	Definitions
$v$	Subscript indicating the visual part of an entity
$t$	Subscript indicating the textual part of an entity
$q = (q_v, q_t)$	Multimedia query (which reduces to $q = (q_t)$ in the asymmetric search scenario)
$d = (d_v, d_t)$	Multimedia object in the database
$n$	The number of multimedia objects or documents in the database
$s_v(q, \cdot)$	Visual similarities (row) vector of $q$ with all documents of the database (of size $1 \times n$ )
$s_t(q, \cdot)$	Textual similarities (row) vector of $q$ (of size $1 \times n$ )
$l$	The number of top elements retained from $s_t$ for semantic filtering.
$S_v$	Visual similarity (square) matrix between pairs of documents (of size $n \times n$ )
$S_t$	Textual similarity (square) matrix between pairs of documents (of size $n \times n$ )
$s_*^{q_t}, S_*^{q_t}$	Same as above but text query semantically filtered (of size $1 \times l$ and $l \times l$ )
$\mathbf{K}(\cdot, k)$	$k$ nearest neighbor thresholding operator acting on a vector
$x_{(i)}$	Diffusion process iteration on the full graph, starting from the text modality (of size $1 \times n$ )
$y_{(i)}$	Diffusion process iteration on the full graph, starting from the visual modality (of size $1 \times n$ )
$x_{(i)}^{q_t}, y_{(i)}^{q_t}$	Same as above but using the graph reduced with the $q_t$ based semantic filter (of size $1 \times l$ )
$cm_{tv}^{q_t}, cm_{vt}^{q_t}$	Cross-media similarities corresponding to $x_{(1)}^{q_t}$ respectively to $y_{(1)}^{q_t}$
$rw_{tv}^{q_t}, rw_{vt}^{q_t}$	Random walk based scores corresponding to $x_{(\infty)}^{q_t}$ respectively to $y_{(\infty)}^{q_t}$ , with $k = l$
$gd_{tv}^{q_t}, gd_{vt}^{q_t}$	Generalized diffusion model corresponding to $x_{(\infty)}^{q_t}$ respectively to $y_{(\infty)}^{q_t}$ , with $k \ll l$ .

Table 1: Notations and definitions.

- The  $\cdot$  symbol represents the regular matrix multiplication operation.

The previously introduced cross-media similarity, denoted  $cm_{tv}(q, \cdot)$ , propagates the text similarities of pseudo-relevant objects to their visual similarities which can be seen as a transmedia pseudo-relevance feedback mechanism. This operation is non commutative and we can design a cross-media similarity,  $cm_{vt}(q, \cdot)$ , propagating visual similarities to textual similarities, providing that we are also given an image query  $q_v$ . We then obtain :

$$cm_{vt}(q, \cdot) = \mathbf{K}(s_v(q, \cdot), k) \cdot S_t \quad (2)$$

These cross-media similarities attempt to bridge the semantic gap between visual and textual information by enriching one modality by the other using monomedia nearest neighbors as proxies. Once the cross-media similarities are computed we can linearly combine them with monomedia similarities as follows :

$$rsv_{cm}(q, \cdot) = \alpha_t s_t(q, \cdot) + \alpha_v s_v(q, \cdot) + \alpha_{tv} cm_{tv}(q, \cdot) + \alpha_{vt} cm_{vt}(q, \cdot) \quad (3)$$

where  $\alpha_t, \alpha_v, \alpha_{tv}, \alpha_{vt}$  are real parameters that sum to one.

The formula given in Eq. 3 encompasses different particular sub-cases :

- $\alpha_{tv} = \alpha_{vt} = 0$ , leads to the classic late fusion technique using a weighted mean as an aggregation function.
- $\alpha_v = \alpha_{vt} = 0$ , gives a cross-media based approach to address CBMIR tasks in the context of the asymmetric case.
- $\alpha_v = \alpha_{tv} = 0$ , is one particular combination that gave top-ranked results on different ImageCLEF tasks [15, 2, 3]. Indeed, it was already shown that the visual information is particularly beneficial through the cross-media  $cm_{vt}$  scores that spread visual information to textual information.

Cross-media similarities draw inspiration from Cross-Media Relevance Models [32] and intermedia feedback methods proposed in [42]. For example, from an image query, a first visual similarity is computed and an initial set of (assumed) relevant objects is retrieved. As the objects are multimodal, each image has also a text part, and this text can feed any text feedback method (other than relevance models). In other words, the modality of data is switched, from image to text or text to image, during the (pseudo) feedback process. In that sense, cross-media techniques generalize the pseudo-feedback idea present in the cross-media relevance model.

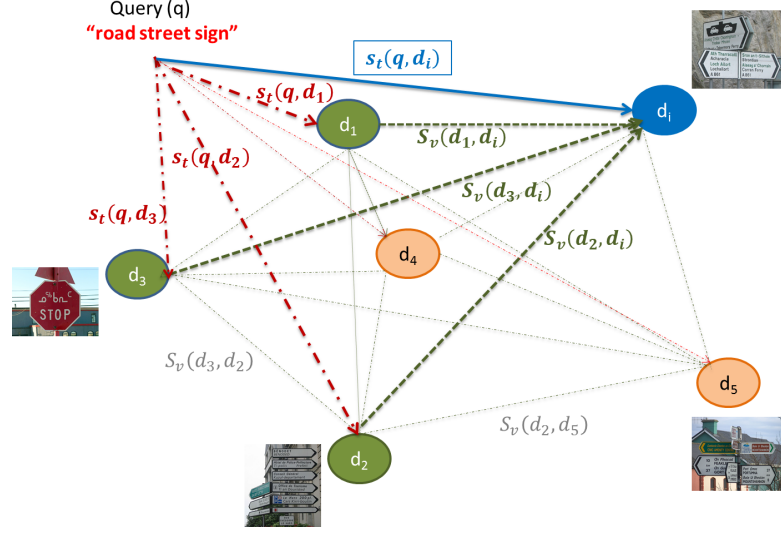


Figure 3: Given a text query  $q_t$  the cross-media relevance score can be computed as  $\sum_{d_j \in \mathcal{N}_t(q)} s_t(q, d_j) \cdot S_v(d_j, d_i)$ . Note that the sum is over the nearest neighbors of the “query”, hence the complementary visual information of the documents that are close to the query are exploited.

### 3.2 Methods based on random walks

The PageRank algorithm proposed in [9, 36, 26] has been an important step forward in development and success of search engines such as Google. It is therefore not surprising that multimedia information fusion based on graph modeling using random walks has been addressed by several researchers [47, 31, 59, 40]. In this paper we particularly study the method proposed in [31, 30]. In this approach, it is assumed that each image is a node of a graph and two images are linked with a weighted edge if there exists a multimodal contextual similarity between them ((see also Figure 4). Depending on the application, the definition of such multimodal contextual similarities can vary. Typically, we assume that they are given by a linear combination of some visual and textual similarities.

The research work described in [31] deals with video retrieval. In the latter paper, the authors propose to use near-duplicate detection measures as for visual similarities between video stories. Text similarities are derived from automatic speech recognition and machine translation transcripts and measured by a mutual information approach.

In our perspective, we are concerned with image/text data and we assume generic image based and text based similarity matrices which are respectively denoted  $S_v$  and  $S_t$ . Using the notations given in Table 1, the multimodal contextual similarity matrix according to [31], that we denote by  $C$ , can be interpreted as follows :

$$C = (1 - \beta)S_v + \beta S_t \quad (4)$$

where  $\beta \in [0, 1]$ .

We then transform  $C$  into a stochastic matrix, denoted by  $P$ , by applying the following normalization operator<sup>2</sup> :

$$P = D \cdot C \quad (5)$$

where  $D$  is the diagonal matrix of size  $n \times n$ , with general term  $D(i, i) = 1 / \sum_{j=1}^n C(i, j)$  and  $D(i, j) = 0$  for all  $i \neq j$ .

The general term  $P(i, j)$  is interpreted as the probability to go from “state”  $i$  to “state”  $j$  where these indices respectively refer to documents  $d^i$  and  $d^j$ . We then compute the random walk’s stationary probability distribution over the documents. Such graph based measures are then employed to rerank the list retrieved by the text based

<sup>2</sup>Note that before normalization,  $P$  can be sparsified, *i.e.* only top pairwise similarities are considered for each document as shown in Figure 4. When  $P$  is fully computed, it means that the context of each document is the whole dataset.



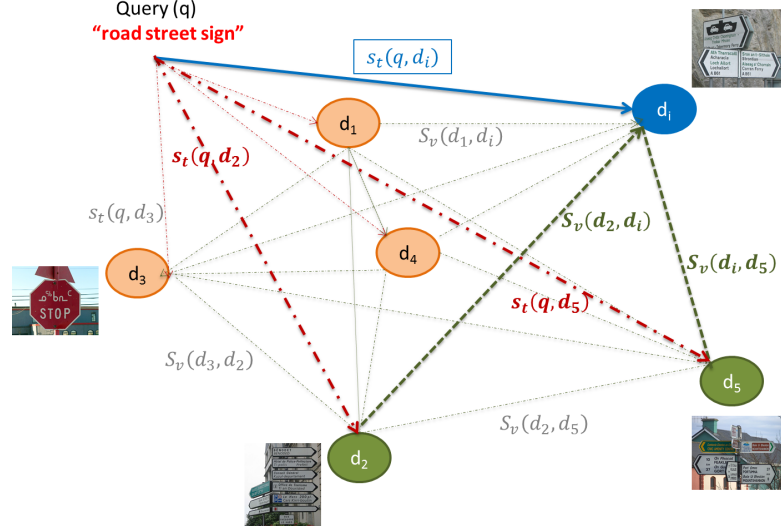


Figure 4: Given a text query  $q_t$  the new relevance score can be computed as  $s_t(q, d_i) + s_t(q, d_i) + \sum_{d_j \in \mathcal{N}_v(d_i)} s_t(q, d_j) \cdot S_v(d_j, d_i)$ . Note that in contrast to the cross-media, the sum is over the nearest neighbors of the document  $d_i$ , hence the visual (or multi-modal if  $P = (1 - \beta)S_v + \beta S_t$  are considered) context of the document  $d_i$  is exploited.

scores. However, to further fuse visual and textual information, the random walk is biased towards documents with higher textual similarity values with the text query. In other words, we add a prior based on the text scores in the random walk process. Note that such a prior can also be interpreted as a restart process or a personalization vector in other information retrieval tasks.

Formally, if we denote by  $x_{(i)}$  the row vector of size  $1 \times n$  of the state probabilities at iteration  $i$  then we have :

$$x_{(i)} = (1 - \gamma)x_{(i-1)} \cdot P + \gamma s_t(q, \cdot) \quad (6)$$

where  $\gamma \in [0, 1]$ .

In order to obtain the state stationary distribution, we iterate the previous updating equation until convergence which yields to the following definition :

$$x_\infty = (1 - \gamma)x_\infty \cdot P + \gamma s_t(q, \cdot) \quad (7)$$

In [31] only the asymmetric search scenario with a text query solely was treated. In this paper, we consider the extension of this approach when we are also given an image query. Accordingly, we use a similar random walk process but with a prior depending on the initial image based scores  $s_v(q, \cdot)$  and define the related stationary distribution :

$$y_\infty = (1 - \gamma)y_\infty \cdot P + \gamma s_v(q, \cdot) \quad (8)$$

Let us denote  $rw_{tv}(q, \cdot) = x_\infty$  and  $rw_{vt}(q, \cdot) = y_\infty$ . We can linearly combine these graph based scores with the initial monomedia similarities and design the following final relevance score :

$$rsv_{rw}(q, \cdot) = \alpha_t s_t(q, \cdot) + \alpha_v s_v(q, \cdot) + \alpha_{tv} rw_{tv}(q, \cdot) + \alpha_{vt} rw_{vt}(q, \cdot) \quad (9)$$

We can consider the following particular cases :

- $\alpha_{vt} = \alpha_{tv} = 0$ , leads to the classic late fusion technique as for cross-media similarities.
- $\alpha_t = \alpha_v = \alpha_{vt} = 0$ , is a combination that reduces to  $rw_{tv}$ . It assumes the asymmetric search scenario and was tested<sup>3</sup> in [31].

<sup>3</sup>This combination was named FRTP in [31]



Figure 5: Top retrieved images with CBIR and CBMIR for the topic 22 at ImageCLEF Wikipedia Challenge 2010.

- $\alpha_v, \alpha_{vt} > 0$ , is, to our knowledge, a new extension of the method which assumes the symmetric search scenario.

Before analyzing further the two graph methods we have introduced, we discuss in the sequel, an important aspect of the combination of visual and textual information in CBMIR. Our multimedia retrieval model, we are going to introduce in section 5, results from the materials described both in the present and the next sections.

## 4 Text query based semantic filtering of multimedia similarities

We first underline the particular importance of textual similarities between the text query  $q_t$  and the text part of multimedia items of the database when addressing CBMIR tasks. Our observations lead us to propose the text based semantic filtering of multimedia similarities that we argue to be a crucial pre-processing step in CBMIR and thus in our multimedia retrieval model. As we shall see, this approach is similar in spirit to the visual reranking paradigm. However, in our perspective, we generalize the latter concept by applying the semantic filtering not only to the visual similarities between the query and the documents in the collection but also to any similarities we employ in our fusion model, such as visual or textual similarities between the documents in the collection. Before formally stating the text query based semantic filtering method, we provide the rationale of such an approach by discussing the semantics conveyed by textual similarities as compared to visual similarities.

When text is used as query, only a few keywords are usually provided. In contrast, when an image is used as query, "all the information it contains" is provided to the system. It is generally said that "a picture is worth a thousand words" but in the context of information retrieval, which word(s) is meant when an image is used as a query? Content based image retrieval (CBIR) systems attempt to find similar images of an image query from a visual standpoint but in most cases the user is rather interested in some underlying semantic meanings of the image query.

To illustrate this ambiguity, let us consider the example given in Figure 5. With regard to this topic, we provide the results obtained with a CBIR system which uses an image query and also the ones provided by a CBMIR system which uses both visual and textual information. If we only use the image query, we can see that the CBIR system retrieves visually similar images but these latter items are in fact irrelevant to the information need of the user. Indeed, the text query associated with this topic is "shark underwater". The images retrieved by the CBIR system show a blue background with, most of them, objects with a fish-like shape. However, none of these images contains a shark. There is a semantic mismatch between the user's information need and the images retrieved by the CBIR system. On the contrary, the images retrieved by the CBMIR system contain sharks for most of them even if their visual similarity values with the image query are lower than the ones given by the CBIR system. As a result, the list retrieved by the CBMIR system is more relevant because the text query gives the semantic meaning of the images the user is interested in unlike the CBIR system.

In this paper, we apply the semantic filtering method described in [16]. This filtering operator aims at semantically correcting visual similarities between two multimedia items by using their textual similarities and this approach already showed to lead to better results in CBMIR. The semantic operator defined in [16] amounts to

filtering visual scores as follows :

$$s_v^{qt}(q, d) = \begin{cases} s_v(q, d) & \text{if } d \text{ is in the top } l \text{ list according to } s_t(q, \cdot) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

The number of selected documents we consider (the nearest neighbors) is equal to  $l = \min(\text{nnz}(s_t(q, \cdot)), m)$ , where  $\text{nnz}(s_t(q, \cdot))$  corresponds to the number of non zero textual scores given the query  $q$ , and  $m$  is a maximum number of documents which are considered to be pseudo-relevant ( $m = 1,000$  in our experiments).

If image reranking [6, 51] implicitly uses a similar approach, filtering is essentially thought of as a pragmatic way to combine text and image scores, as the text is only used to select the documents to be ranked and the ranking is done by the visual scores. Here, we adopt the view of [16] where the filtering step similarly used to select the documents to be considered, but after the visual scores are recombined with the textual scores for a final ranking. This fundamental difference lead [16] to the following method : after having semantically filtered the visual similarities  $s_v(q, \cdot)$  in order to obtain  $s_v^{qt}(q, \cdot)$ , a late fusion approach scheme based upon the weighted mean between  $s_v^{qt}(q, \cdot)$  and  $s_t(q, \cdot)$  was used to rank and it was showed that such an approach outperforms other late fusion methods and also the basic visual reranking method<sup>4</sup>. Similarly, here we use the filtering to pre-select the documents on which we apply our relevance models. We argue that there is an inherent ambiguity when one use visual query and filtering is thought of as a way to correct visual similarities and to specify an information need. We admit nevertheless that this approach has the limitation of ignoring (loosing) relevant visual documents with no textual or irrelevant textual information. Newertheless, note that our textual filtering step uses text retrieval techniques that goes beyond simple keyword matching (see appendix A) and is able to retrieve documents that have semantic similarity with the query (using *e.g.* lexical entailment and query expansion).

Hence, in this paper, we extend [16] by using the semantic filtering strategy in the context of graph based methods. We propose to apply this filtering scheme to any multimedia scores and similarities before employing a graph based relevance model. We thus use the top  $l$  list given by  $s_t(q, \cdot)$  to semantically filter all other similarity matrices and relevance scores. Indeed, we have previously argued that text based relevance scores are usually better in retrieving relevant documents in CBMIR. Therefore, we want to favor the top  $l$  list given by  $s_t(q, \cdot)$  in any similarities and relevance scores involved in the fusion process. To this end, we apply the same kind of semantic filter given in Eq. 10 not only to  $s_v(q, \cdot)$  but also to  $S_v$ ,  $S_t$  and  $s_t(q, \cdot)$  itself. As a consequence, we introduce the following text query based **semantically filtered visual similarities**<sup>5</sup> :

$$S_v^{qt}(d, d') = \begin{cases} S_v(d, d') & \text{if } d \text{ and } d' \text{ are in the top } l \text{ list according to } s_t(q, \cdot) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Similarly, we respectively define text query based semantically filtered  $s_t^{qt}(q, \cdot)$  and  $S_t^{qt}$  as follows :

$$s_t^{qt}(q, d) = \begin{cases} s_t(q, d) & \text{if } d \text{ is in the top } l \text{ list according to } s_t(q, \cdot) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$S_t^{qt}(d, d') = \begin{cases} S_t(d, d') & \text{if } d \text{ and } d' \text{ are in the top } l \text{ list according to } s_t(q, \cdot) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Therefore, in what follows, the vectors  $s_t^{qt}(q, \cdot)$  and  $s_v^{qt}(q, \cdot)$  are sparse and contain only  $l \ll n$  non zero elements. Similarly, each row of the matrices  $S_t^{qt}$  and  $S_v^{qt}$  only contains  $l$  non zero<sup>6</sup>

Moreover, since we are only interested in the top  $l$  list provided by the text relevance scores, we can also remove from  $S_t^{qt}$  and  $S_v^{qt}$  the rows and columns of items that do not belong to this list. As a consequence, in

<sup>4</sup>Note that the method proposed in [16] amounts to linearly combining the visual reranking scores and the textual scores, which makes this method different from the visual reranking techniques presented in paragraph 2.2. Indeed in the latter case the top selected documents are ranked based on their visual similarity, while the former ranks the documents based on the fused scores and has been shown to yield a much better retrieval performance.

<sup>5</sup>Note that while the retrieval process using such similarities is not any more pure visual retrieval, the similarities scores themselves  $S_v^{qt}(d, d')$  are purely visual similarities computed between the visual signatures of  $d$  and  $d'$ .

<sup>6</sup>The main idea is that given a query the top  $l$  documents are selected from the collection and the rest of the collection is not considered. In terms of matrix representation  $S_v^{qt}$  is a sparsification of  $S_v$  where elements  $(i, j)$  of the matrix are set to zero except the ones where both  $d_i$  and  $d_j$  are amongst the top  $l$  selected ones. The non-zero elements of  $S_v^{qt}$  form an  $l \times l$  sub-matrix of  $S_v$ . In the case of the text the aim is less the semantic alignment but the sparsification makes the computational cost feasible for large scale datasets.

practice, in order to alleviate the memory and time complexities of graph based techniques, when we compute the relevance and similarity values with respect to a text query  $q_t$ , we consider  $s_t^{q_t}(q, \cdot)$  and  $s_v^{q_t}(q, \cdot)$  as vectors of size  $1 \times l$  and  $S_t$  and  $S_v$  as matrices of size  $l \times l$ .

Consequently, the semantic filtering approach not only allows one to better bridge the semantic gap but it also dramatically improves the memory complexity since we only need to store matrices of size  $O(l^2)$  instead of  $O(n^2)$ . Furthermore, as we discussed in the previous section, graph based methods rely on diffusion processes which, from an algebraic viewpoint, are materialized by matrix multiplication operations. Since this calculation has a cubic computation complexity with respect to the size of the matrix, the semantically filtered technique enables reducing the time complexity of the graph based methods as well, notably from  $O(n^3)$  to  $O(l^3)$ . Overall, this method makes the graph based techniques scalable for very large multimedia repositories. From a more theoretical standpoint, we introduce in the sequel our multimedia relevance model which makes use of the text query based semantic filtering as a core principle and which relies on an unifying framework for graph based techniques that encompasses the methods we have detailed in section 3.

## 5 A unifying framework using semantic filtering and graph based methods

We now introduce our multimedia retrieval model. Firstly, our framework uses the text query based semantic filtering as a first level of information fusion. In other words, given a text query  $q_t$ , we start by selecting a subset of semantically relevant items by restraining the search space to the top  $l$  elements provided by the textual scores  $s_t(q, \cdot)$ . In practice, we apply the semantic filters given by Eqs. 10, 11, 12 and 13. We then propose to apply the graph based methods described in section 3 to the text query based semantically filtered scores and similarities. This leads us to define the following cross-media similarities and derived relevance scores :

$$cm_{tv}^{q_t}(q, \cdot) = \mathbf{K}(s_t^{q_t}(q, \cdot), k) \cdot S_v^{q_t} \quad (14)$$

$$cm_{vt}^{q_t}(q, \cdot) = \mathbf{K}(s_v^{q_t}(q, \cdot), k) \cdot S_t^{q_t} \quad (15)$$

As argued previously in section 3 we can further fuse the cross-media similarities with semantically filtered textual and visual scores :

$$rv_{cm}^{q_t}(q, \cdot) = \alpha_t s_t^{q_t}(q, \cdot) + \alpha_v s_v^{q_t}(q, \cdot) + \alpha_{tv} cm_{tv}^{q_t}(q, \cdot) + \alpha_{vt} cm_{vt}^{q_t}(q, \cdot) \quad (16)$$

In the same manner, we can define retrieval models based on random walks over the stochastic matrix obtained from the semantically filtered multimedia similarities. For the method that takes into account a text based prior given by  $s_t^{q_t}(q, \cdot)$ , we have :

$$x_{(i)}^{q_t} = (1 - \gamma) x_{(i-1)}^{q_t} \cdot P^{q_t} + \gamma s_t^{q_t}(q, \cdot) \quad (17)$$

where  $P^{q_t} = D^{q_t} \cdot C^{q_t}$ ,  $C^{q_t} = (1 - \beta) S_v^{q_t} + \beta S_t^{q_t}$  and  $D^{q_t}$  is the diagonal matrix whose entries are such that  $D^{q_t}(i, i) = 1 / \sum_{j=1}^n C^{q_t}(i, j)$  and  $D^{q_t}(i, j) = 0$  for all  $i \neq j$ . The stationary distribution of the previous equation is such that<sup>7</sup> :

$$x_{\infty}^{q_t} = (1 - \gamma) x_{\infty}^{q_t} \cdot P^{q_t} + \gamma s_t^{q_t}(q, \cdot) \quad (18)$$

Then, for the random walk based scores with a prior depending on  $s_v^{q_t}(q, \cdot)$ , its stationary distribution is given by :

$$y_{\infty}^{q_t} = (1 - \gamma) y_{\infty}^{q_t} \cdot P^{q_t} + \gamma s_v^{q_t}(q, \cdot) \quad (19)$$

Following Eq. 9, we can further linearly fused the random walk based scores with  $s_t^{q_t}(q, \cdot)$  and  $s_v^{q_t}(q, \cdot)$  which yields to :

$$rv_{rw}^{q_t}(q, \cdot) = \alpha_t s_t^{q_t}(q, \cdot) + \alpha_v s_v^{q_t}(q, \cdot) + \alpha_{tv} rv_{tv}^{q_t}(q, \cdot) + \alpha_{vt} rv_{vt}^{q_t}(q, \cdot) \quad (20)$$

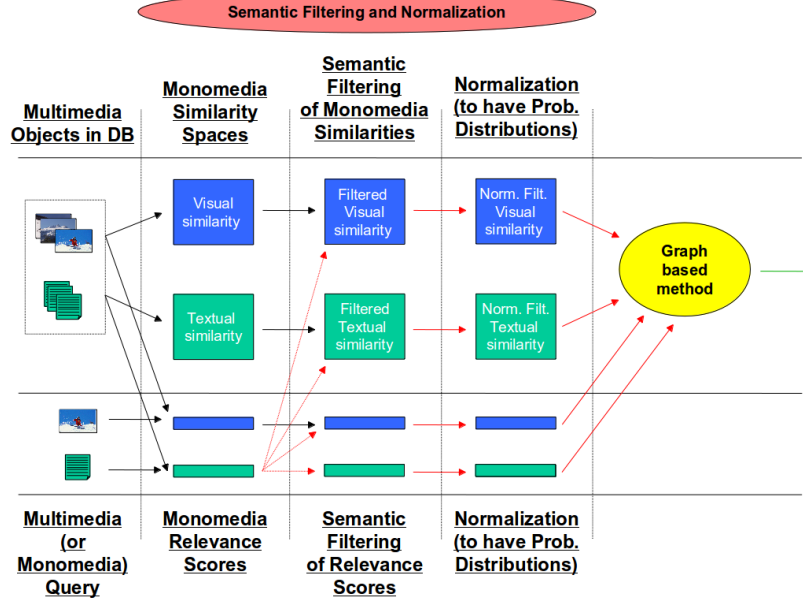


Figure 6: Pre-processing, text query based semantic filtering and normalization.

In Figure 6, we depict the first feature of our multimedia retrieval model which applies the text query based semantic filtering to the query and to the multimedia items of the database.

The second feature of our multimedia retrieval model aims at defining a unifying framework for graph based methods that encompasses the diffusion processes strategies underlying both the cross-media similarities and the random walk based scores. Note that in order to embed these two approaches in the same model, we assume that all similarities have been normalized so that we manipulate probability distributions. Henceforth, we assume that  $s_t^{qt}(q, \cdot)$ ,  $s_v^{qt}(q, \cdot)$ , and rows of  $S_t^{qt}$  and the ones of  $S_v^{qt}$  have non negative values and that they all sum to one<sup>8</sup>. This constraint is due to the random walk method but the cross-media approach does not initially require such a normalization and other possibilities exist. We will come back to this point later on in section 7.3. The normalization step occurs just after the text query based semantic filtering and just before applying graph based methods as depicted in Figure 6.

To establish our unifying graph based model, let us start by studying the random walk approach a little bit deeper and let us consider the following formula :

$$x_\infty^{qt} = (1 - \gamma)x_\infty^{qt} \cdot P^{qt} + \gamma x_\infty^{qt} \cdot e \cdot s_t^{qt}(q, \cdot) \quad (21)$$

where  $e$  is the  $l \times 1$  vector full of 1.

In the previous equation, the sub-part  $x_\infty^{qt} \cdot e$  reduces to 1 since  $x_\infty^{qt}$  is a probability distribution. Therefore Eq. 21 and Eq. 18 are strictly equivalent. But, in Eq. 21, we can factorize the term  $x_\infty^{qt}$  to obtain :

$$x_\infty^{qt} = x_\infty^{qt} \cdot [(1 - \gamma)P^{qt} + \gamma e \cdot s_t^{qt}(q, \cdot)] \quad (22)$$

Let us introduce the following matrix of size  $l \times l$ :

$$Q_{tv}^{qt} = (1 - \gamma)P^{qt} + \gamma e \cdot s_t^{qt}(q, \cdot) \quad (23)$$

Using this matrix, Eq. 22 can be re-written as  $x_\infty^{qt} = x_\infty^{qt} \cdot Q_{tv}^{qt}$ . The solution of this equation is the same as the solution of  $(x_\infty^{qt})^\top = (Q_{tv}^{qt})^\top \cdot (x_\infty^{qt})^\top$  where the right superscript  $\top$  states for the transpose operation on vectors and matrices. From the latter relation we see that the stationary probability distribution of the random

<sup>7</sup>Denoted in [31] by PRTP

<sup>8</sup>This amounts to dividing the row vectors by their  $L1$  norms.

walk is related to an eigen-decomposition problem [36]. Indeed,  $x_\infty^{qt}$  is clearly the eigenvector of  $(Q_{tv}^{qt})^\top$  associated to the eigenvalue 1. Since  $Q_{tv}^{qt}$  is a stochastic matrix, 1 is the highest eigenvalue. As a result,  $x_\infty^{qt}$  is the leading eigenvector of  $(Q_{tv}^{qt})^\top$ . One efficient way to compute the leading eigenvector of a square matrix is the power method [36]. Thus, in practice, we iterate the following equation until convergence in order to determine  $rw_{tv}^{qt}(q, \cdot)$ :

$$(x_{(i)}^{qt})^\top = (Q_{tv}^{qt})^\top \cdot (x_{(i-1)}^{qt})^\top \quad (24)$$

Since  $x_{(0)}^{qt}$  is a probability distribution then so are the vectors  $x_{(i)}^{qt}, i > 0$  and  $x_\infty^{qt}$  represents the stationary distribution which is proportional to the leading eigenvector of  $Q_{tv}^{qt}$ .

Let us now consider the following general formula :

$$\begin{aligned} x_{(i)}^{qt} &\propto \mathbf{K}(x_{(i-1)}^{qt}, k) \cdot [(1 - \gamma)D^{qt} \cdot (\beta S_t^{qt} + (1 - \beta)S_v^{qt}) + \gamma e \cdot s_t^{qt}(q, \cdot)] \\ x_{(0)}^{qt} &= s_t^{qt}(q, \cdot) \end{aligned} \quad (25)$$

$x_{(\cdot)}^{qt}$  can be interpreted as a generalized diffusion process with a text based prior. From the previous development (Eqs. 23 and 24), we can see that  $rw_{tv}^{qt}(q, \cdot)$  can be derived from Eq. 25 given by the limit vector  $x_\infty^{qt}$  when we use  $k = l$  as  $\mathbf{K}(x_{(i-1)}^{qt}, l)$  is equivalent to  $x_{(i-1)}^{qt}$  (no operator  $\mathbf{K}$  is applied). In this case, the random-walk is guaranteed to converge and the limit value does not depend on the initialization.

We can see that the generic case Eq. 25 combines the idea of considering only a few nearest neighbors in the diffusion process as in the case of the cross-media, while doing several iterations (until stability) as in the case of the random walk. To avoid confusion, we will refer to this approach as the generalized diffusion model and denote it by  $gd_{tv}^{qt}$ .

When  $k < l$ , we do not have a theoretical guarantee of the convergence of the diffusion process. However, we have experimentally observed that after several iterations the scores became stable. It seems that the set of top  $k$  documents remains unchanged throughout the different iterations. In this case, Eq. 25 becomes quasi-equivalent to a power iteration as formalized by Eq. 24 but with the corresponding reduced graph of size  $k \times k$ . Indeed, we can see that the zeros in  $\mathbf{K}(x_{(i-1)}^{qt}, k)$  will eliminate from  $Q_{tv}^{qt}$ , the rows corresponding to the documents not selected by the operator  $\mathbf{K}$ . Concerning the columns corresponding to these documents, while they contribute to create  $x_{(i)}^{qt}$ , the scores corresponding to them will be ignored in the next step when we will apply the operator  $\mathbf{K}$  on  $x_{(i)}^{qt}$  (note that we assumed that the set of top  $k$  documents do not change any more in the iterations).

Note that the generalization of the random walk process with the  $\mathbf{K}$  operator is new in the literature and we are not aware of any similar work. It is indeed different from the Link Reduction by  $k$  nearest neighbors (PRTP-KNN) proposed in [31], where the  $k$  nearest neighbors are considered for each node in the graph. This can be seen as a sparsification of the matrix  $S^{qt}$  where in each row and column only the  $k$  highest values are kept non-zero. We did not consider and tested such sparsification in our experiments, as the best PRTP-KNN yields the same results as PRTP.

On the other hand, let us now consider  $\gamma = 0$ , which cancels the prior given by the text based scores;  $\beta = 0$ , which cancels the text based similarity matrix in the convex combination in Eq. 17; and let us iterate Eq. 25 only once ( $i = 1$ ). In this particular setting  $x_{(1)}^{qt}$  actually corresponds to the cross-media similarity  $cm_{tv}^{qt}(q, \cdot)$  given in Eq. 14.

From these previous observations, we have shown that Eq. 25 is a general graph based approach which generalizes both  $rw_{tv}^{qt}(q, \cdot)$  and  $cm_{tv}^{qt}(q, \cdot)$  methods.

Similarly, we propose the following formula that allows us to generalize the symmetric relations  $rw_{vt}^{qt}(q, \cdot)$  and  $cm_{vt}^{qt}(q, \cdot)$ :

$$\begin{aligned} y_{(i)}^{qt} &\propto \mathbf{K}(y_{(i-1)}^{qt}, k) \cdot [(1 - \gamma)D^{qt} \cdot (\beta S_v^{qt} + (1 - \beta)S_t^{qt}) + \gamma e \cdot s_v^{qt}(q, \cdot)] \\ y_{(0)}^{qt} &= s_v^{qt}(q, \cdot) \end{aligned} \quad (26)$$

In that case,  $y_{(\cdot)}^{qt}$  is a generalized diffusion process with a semantically filtered image based prior. In the right member of Eq. 26, what formally changes as compared to Eq. 25, is the substitution of  $t$  by  $v$  and *vice versa*. However, as stated in the introduction, this formula allows one to consider the symmetric search scenario that has been less investigated in the context of the random walk approach. In such a case, we suppose not only a text query but also an image query and we can thus consider using the random walk technique for multimedia fusion

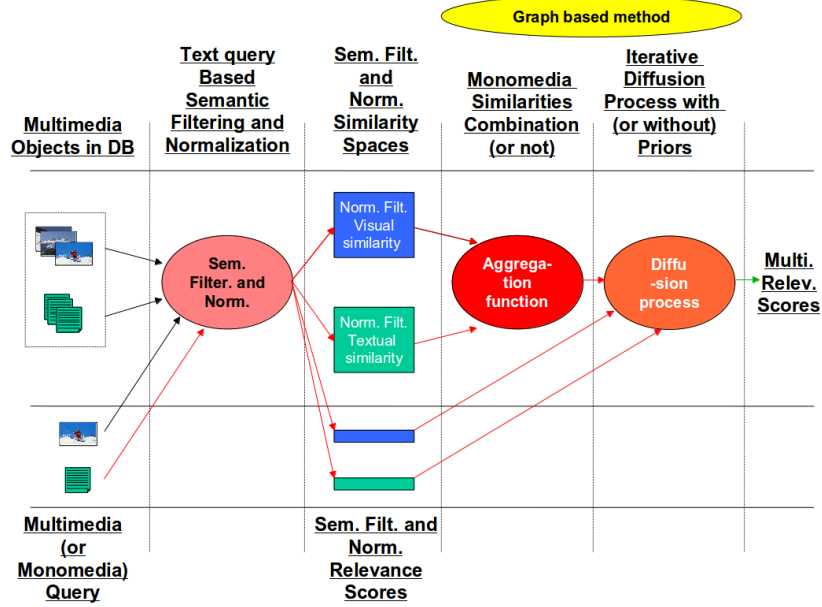


Figure 7: Unified view of graph based methods and our multimedia retrieval model.

using a semantically filtered visual based prior. Indeed in Eq. 26, we obtain a random walk based technique biased towards  $s_v^{q_t}(q, \cdot)$  that will converge to  $rw_{vt}^{q_t}(q, \cdot) = y_{\infty}^{q_t}$  given by Eq. 19.

As far as the cross-media based approach is concerned, we obtain the already defined  $cm_{vt}^{q_t}(q, \cdot)$  in Eq. 15 from the Eq. 26 by using  $\gamma = 0$ , which cancels the prior given by the image based scores;  $\beta = 1$ , which cancels the image based similarity matrix in the convex combination and we iterate Eq. 26 only once ( $cm_{vt}^{q_t}(q, \cdot) = y_{(1)}^{q_t}$ ).

This unifying framework encompasses the cross-media similarities and the random walk based method for CBMIR. Eq. 25 and Eq. 26 allow us to have a better understanding of the main differences between these two techniques from a conceptual point of view. However, our proposal suggests more than a simple comparison of those two approaches, it invites to a deeper analysis of what are the key points when using graph based techniques in CBMIR.

We depict in Figure 7 the unified formulation of graph based approaches that we have introduced previously accompanied with the preliminary semantic filtering and normalization steps. Overall, this schema represents the multimedia retrieval model we propose in this paper.

In the following sections, we experiment with the proposed multimedia retrieval model in the case of content based image/text multimedia retrieval. We will particularly focus on the comparison of three orientations of our framework :

- the one that leads to cross-media similarities :  $cm_{tv}^{q_t} = x_{(1)}^{q_t}$  and  $cm_{vt}^{q_t} = y_{(1)}^{q_t}$ ;
- the one that reduces to random walk based scores :  $rw_{tv}^{q_t} = y_{(\infty)}^{q_t}$ ,  $rw_{vt}^{q_t} = y_{(\infty)}^{q_t}$  and  $k = l$ , meaning that we do not use the operator  $\mathbf{K}$ ;
- and the generalized diffusion model :  $gd_{tv}^{q_t} = y_{(\infty)}^{q_t}$ ,  $gd_{vt}^{q_t} = y_{(\infty)}^{q_t}$  and  $k \ll l$ .

## 6 Experimental settings

Firstly, we describe the real-world datasets we applied the different tested techniques to. Then, we introduce the image and text representations and similarities we used in our experiments.

## 6.1 Datasets

We conducted our experiments on real-world collections which are constituted of image/text items. The first two datasets were used in the ImageCLEF Photo or Wikipedia retrieval tasks<sup>9</sup> while the last one was constituted in order to assess web image search techniques<sup>10</sup>. We give below the description of these repositories and the tasks they were meant to address, according to the respective websites that present them.

- The IAPR dataset was used in the context of ImageCLEF 2008 [28]. “The image collection of the IAPR TC-12 Benchmark consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Each image is associated with a text caption in up to three different languages (English, German and Spanish) . These annotations are stored in a database which is managed by a benchmark administration system that allows the specification of parameters according to which different subsets of the image collection can be generated.”
- The Wikipedia collections WIKI10 and WIKI11 were used in ImageCLEF 2010 and 2011 [50]. “The Wikipedia image retrieval task is an ad-hoc image retrieval task. The overall goal of the task is to investigate how well multi-modal image retrieval approaches that combine textual and visual evidence in order to satisfy a user’s multimedia information need could deal with larger scale image collections that contain highly heterogeneous items both in terms of their textual descriptions and their visual content. The aim is to simulate image retrieval in a realistic setting, such as the Web environment, where available images cover highly diverse subjects and have highly varied visual properties, while their accompanying textual metadata (if any) are user-generated and correspond to noisy and unstructured textual descriptions of varying quality and length.”<sup>11</sup>. Both collections actually contain the same set of 237,434 images. The difference between WIKI10 and WIKI11 is the set of topics used in order to take into account several kinds of multimedia information needs. WIKI10 consists in 70 topics while WIKI11 contains 50 topics. “The ground truth for these topics was created by assuming binary relevance (relevant *vs.* non relevant) and by assessing only the images in the pools created by the retrieved images contained in the runs submitted by the participants each year.”
- The Web Queries (WEBQ) repository was used as a benchmark in order to assess the research work described in [35]. “The Web Queries dataset contains 71,478 images and meta-data retrieved by 353 web queries. For each retrieved image the relevance label is available. The relevance labels are obtained by manual labeling. French query words were used to retrieve the images, but we provide also the English translation.” Unlike the previous tasks, WEBQ contains only text topics. Thus, it is a case of asymmetric search scenario.

Though we use three different collections, our experiments concern four tasks : IAPR, WIKI10, WIKI11 and WEBQ. The tasks are all content based image/text multimedia data retrieval ones. On each topic given in each task, we tested different particular cases of the graph based approach introduced in section 5. A topic consists in an image/text query (except for the WEBQ as explained beforehand) and we were also provided with the binary ground truth (relevant *vs.* non relevant). We used the Mean Average Precision (MAP) in order to compare the obtained rankings and the ground truth in the goal of evaluating the different multimodal fusion techniques. We also computed if the results were statistically different using paired t-test at the 95% confidence level.

## 6.2 Monomodal Representation and Similarities

Standard preprocessing techniques were first applied to the textual part of the documents. After stop-word removal, words were lemmatized and the collection of documents indexed with Lemur<sup>12</sup>. We used a standard Dirichlet language model on IAPR and the Lexical Entailment (LE) information retrieval model [20] on the

---

<sup>9</sup><http://www.imageclef.org/datasets>

<sup>10</sup><http://lear.inrialpes.fr/~krapac/webqueries/webqueries.html>

<sup>11</sup>This is the description of the dataset as provided at <http://www.imageclef.org/wikidata>

<sup>12</sup><http://www.lemurproject.org/>



Wikipedia datasets. These models were chosen to remain consistent with our previously published and state-of-the-art results [4, 2, 5, 24, 17]. In fact, the LE model clearly outperforms standard IR models and give a relative improvement of 15% MAP<sup>13</sup>. Note that the LE retrieval model is briefly introduced in appendix section A and was recently rediscovered in [33].

As for image representations, we used the Fisher Vector (FV), proposed in [48], an extension of the popular Bag-of-Visual word (BOV) image representation [56, 22], where an image is described by a histogram of quantized local features. In a nutshell, the Fisher vector consists in modeling the distribution of patches in any image with a Gaussian mixture model (GMM) and then in describing an image by its deviation from this average probability distribution. In a recent evaluation [13], it has been shown experimentally that the Fisher vector was the state-of-the-art representation for image classification. The Fisher Vector approach is described in appendix section B.

For the purpose of this paper, the choices of a particular textual and visual similarity are not of first importance. Our framework only requires as input a text ranking expert and a visual ranking expert. So, any textual/visual approaches could be employed and this is why we have moved the descriptions of our experts in the appendix. Our focus here is on the combination of visual and textual modalities. In fact, we did some preliminary experiments varying the textual and/or the visual features but the behavior concerning the combination and the conclusions we could draw were the same as for the monomodal experts used in the paper. Therefore, they do not bring new insights in our experiments and this is why we did not include these results in this paper.

## 7 Experimental results

This section contains an extended empirical analysis of the differences between the two graph based methods we are interested in. But in a more general perspective, the experiments we conducted aim at studying the different settings one could apply using the generalization we propose in Eq. 25 and Eq. 26. For convenience, we remind these two principal graph based formulas below :

$$\begin{aligned} x_{(i)}^{qt} &\propto \mathbf{K}(x_{(i-1)}^{qt}, k) \cdot [(1 - \gamma)D^{qt}(\beta S_t^{qt} + (1 - \beta)S_v^{qt}) + \gamma e \cdot s_t^{qt}]; & x_{(0)}^{qt} &= s_t^{qt} \\ y_{(i)}^{qt} &\propto \mathbf{K}(y_{(i-1)}^{qt}, k) \cdot [(1 - \gamma)D^{qt}(\beta S_v^{qt} + (1 - \beta)S_t^{qt}) + \gamma e \cdot s_v^{qt}]; & y_{(0)}^{qt} &= s_v^{qt} \end{aligned}$$

Our goal is to establish some guidelines on the combination of visual and textual information in CBMIR using graph based methods. To this end, we study several settings of the previously recalled equations and we particularly pay attention to the ones that allow a meaningful comparison between cross-media similarities and random walk based scores.

Accordingly, using Eq. 25 and Eq. 26, we first examine the impact of several parameters on the cross-media and random walk method :

- What is a good initialization for the graph based methods ?
- Is it beneficial to iterate the power method until convergence ?
- What is the impact of the thresholding operator  $\mathbf{K}$  ?
- In which conditions is it beneficial to integrate a text based or an image based prior in the power method ?

Secondly, we investigate on the late combination of the text query based semantically filtered multimedia scores with the graph based scores given by  $cm_{tv}^{qt}$  and  $cm_{vt}^{qt}$  on the one hand, and the random walk based measures  $rw_{tv}^{qt}$  and  $rw_{vt}^{qt}$  on the other hand. In that perspective, we address the following questions :

- Can we expect benefits from a multimedia query as compared to a text only query ?
- Is it beneficial to linearly combine the initial semantically filtered scores with the ones provided by  $cm^{qt}$ ,  $rw^{qt}$  or  $gd^{qt}$  ?
- In which conditions is it beneficial to proceed to a late fusion of similarity matrices before the power method ?

---

<sup>13</sup>roughly a raw 4% in MAP

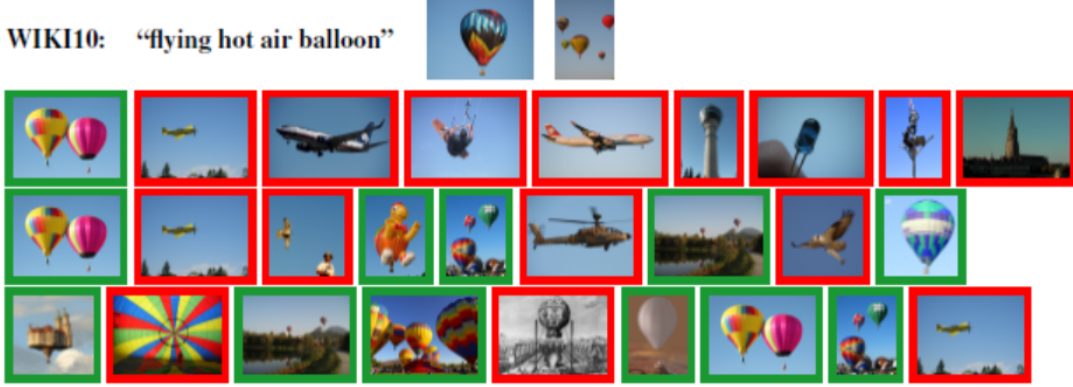


Figure 8: Top retrieved images with pure visual similarity (second row), with semantically filtered visual similarity  $s_v^{qt}$  (third row) and with cross-media  $y_{(i)}$  using  $k = 10$ ,  $\gamma = 0$  and  $i = 1$  (last row), for the topic 9 at ImageCLEF Wikipedia Challenge 2010 (shown in first row). Green means relevant, red non-relevant. Note that the first two “non-relevant” images in the last row are “non-flying” hot air balloons.

All along these empirical analysis, we also comment on the comparison between the asymmetric and symmetric search scenarios. We recall that in the first case, only a text query is assumed in order to search the multimedia collection while in the second case, the user can give an image query in addition to the text query to better express her information need.

Before answering these questions, let us first illustrate three retrieval models for a single query in Figure 8. This figure shows the beneficial effects of : a) the text based semantic filtering and b) the advocated cross-media method.

## 7.1 Comparison of cross-media similarities and random walk based scores

First of all, let us recall that  $x_{(i)}^{qt}$  given in Eq. 25 and  $y_{(i)}^{qt}$  given in Eq. 26 are diffusion processes with priors  $s_t^{qt}$  and  $s_v^{qt}$  respectively. As for the initialization of these iterative equations,  $x_{(0)}^{qt}$  and  $y_{(0)}^{qt}$  could be typically set to uniform distributions. However, some preliminary results showed that such uniform distributions are suboptimal in the case of cross-media and does not affect the classical random walk (without the operator  $\mathbf{K}$ ). As explained previously to encompass the cross-media case we consider as initial distributions  $s_t^{qt}$  and  $s_v^{qt}$  respectively and normalize them to obtain probability distributions.

### 7.1.1 Impact of the number of iterations $i$

In our first set of experiments, we vary the number of iterations  $i$  in Eq. 25 and Eq. 26 using the following setting :

- $\gamma = 0$  (no prior)
- $\beta = 0$  (no late fusion of similarity matrices)

In Table 2 we show the MAP results we obtained when we set  $k = l$  (no nearest neighbor operator). In contrast, in Table 3, the evaluation measures are shown for the best  $k$  among  $\{1, \dots, l\}$ . More precisely, for each task, we first look at the value of  $k$  that provided the best MAP measure after the first iteration and we then iterated the graph based formulas until convergence with this particular value. Best  $k$ , denoted  $k^*$ , were in a rather small range (between 10 and 50) for all tasks (except for WEBQ).

Before focusing on the comparison between cross-media and random walk based results, let us make some preliminary comments :

	IAPR		WIKI10		WIKI11		WEBQ
	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_t^{qt}$
	27.6	26.3	24	26.3	18	27.8	57
$i$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$x_{(i)}^{qt}$
1	<b>28.7</b>	<b>20.8</b>	<b>18.9</b>	<b>15.7</b>	<b>12.6</b>	<b>6.9</b>	69.3
2	23.4 <sup>†</sup>	17.5 <sup>†</sup>	17.2 <sup>†</sup>	13.8 <sup>†</sup>	11.4 <sup>†</sup>	5.3 <sup>†</sup>	<b>69.5<sup>†</sup></b>
3	21.1	15.1	17	13.5	11.3	5.2	68.7
4	19.4	13.4	17	13.5	11.3	5.1	68.4
5	18.7	12.3	17	13.5	11.3	5.1	68.4
10	16.8	9.6	17	13.5	11.3	5.1	68.4
50	15.4	8.5	17	13.5	11.3	5.1	68.4
$\infty$	15.4	8.4	17	13.5	11.3	5.1	68.4

Table 2: **Varying the number of iterations  $i$ .** Results obtained with  $k = l$  (random walk oriented diffusion process) and  $\gamma = \beta = 0$ . The symbol <sup>†</sup> indicates a statistical difference between  $i = 1$  and  $i = 2$  (which implies a statistical difference between  $i = 1$  and  $i > 1$ ).

- From this first set of experiments, we can observe that the graph based scores  $x_{(i)}^{qt}$  and  $y_{(i)}^{qt}$  generally do not outperform the initial scores  $s_t^{qt}$  and  $s_v^{qt}$  with respect to MAP values. In this first step, we indicate that it is not our goal to show that graph based relevance scores outperform the initial semantically filtered visual or textual scores. Our purpose here is rather to compare cross-media based measures against random walk based scores.
- Besides, it is interesting to mention that ranking with the semantically filtered visual relevance scores  $s_v^{qt}$  (given in Table 2 or Table 3), lead to much better results than ranking with pure image scores  $s_v$  (without the semantic filtering) since the latter rankings give 22.1%, 6.2% and 2.7% for tasks IAPR, WIKI10 and WIKI11 respectively<sup>14</sup>. As for the WEBQ task, we only have text based queries. The superiority of  $s_v^{qt}$  over  $s_v$  is particularly true for the Wikipedia repository. It is true that the direct comparison of the ranking based on  $s_v$  which is pure visual with the ranking based on  $s_v^{qt}$  which is multi-modal is not fair. However, it shows that we are improving the MAP performance and hence also the performance on the top, which is primordial as the top documents are used by the operator  $K$  in the trans-modal pseudo-relevance step. Indeed using the textual part of the images in the second row in Figure 8 to enrich the textual part has better chances to improve the results than the texts from the images in the first row.
- Furthermore, when the pure visual scores are reasonably good (such as for the IAPR task), ranking with the semantically filtered visual relevance scores  $s_v^{qt}$  (corresponding to the classical visual reranking method) outperforms the text based relevance scores  $s_t^{qt}$  too. These observations confirm that correcting pure image based similarities using text based similarities is beneficial as stated in section 4. However, as we will see later on, we can further improve the classical visual reranking results by using graph based techniques which provide search results that are complementary.

Let us now analyze Tables 2 and 3 in the goal of comparing the performances of cross-media similarities and random walk based scores. Our first core point concerns the number of iterations these two approaches assume. Indeed, we recall that when  $i = 1$ , the current setting of the parameters of Eq. 25 and Eq. 26 are respectively similar to  $cm_{tv}^{qt}$  and  $cm_{vt}^{qt}$ . In contrast, when the number of iterations  $i$  grows, the graph based relevance scores converge towards  $rw_{tv}^{qt}$  and  $rw_{vt}^{qt}$  which correspond to the case  $i = \infty$ . The results we obtained enable us to claim that going further than a single step in the random walk significantly decreases the performances for all tasks, except for WEBQ, where a second step was beneficial before the system began to degrade. Hence, we conclude that very short walks give better results than walking towards convergence and typically, in our case, a one step walk is the default setting. These results are to be contrasted with the ones obtained in [21], where the authors

<sup>14</sup>Here we refer to the results we obtain when we rank all the documents in the database with the visual scores, i.e. no value in  $s_v$  is set to zero

	IAPR		WIKI10		WIKI11		WEBQ
	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_t^{qt}$
	27.6	26.3	24	26.3	18	27.8	57
$i$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$x_{(i)}^{qt}$
1	<b>35.9</b>	<b>22.4</b>	<b>25.7</b>	<b>23.9</b>	<b>21.4</b>	<b>22.5</b>	69.3
2	32.5 <sup>†</sup>	20.5 <sup>†</sup>	23.9 <sup>†</sup>	21.3 <sup>†</sup>	19.1 <sup>†</sup>	18.3 <sup>†</sup>	<b>69.7<sup>†</sup></b>
3	32.3	19.3	23.4	20.3	18	16	69.1
4	32	18.3	22.8	19.9	18	15	68.8
5	31.9	17.5	22.6	19.6	18.3	14.5	68.8
10	31.7	16.3	22.3	19.2	18.7	14.1	68.8
50	31.7	15.8	22.1	19.1	18.6	14	68.8
100	31.6	15.2	22.1	19.1	18.6	14	68.8
$\infty$	31.6	15.2	22.1	19.1	18.6	14	68.8

Table 3: **Varying the number of iterations  $i$ .** Results obtained with  $k^*$  (best  $k$  obtained for the first step  $i = 1$ ) (cross-media then generalized oriented diffusion processes) and  $\gamma = \beta = 0$ . The symbol <sup>†</sup> indicates a statistical difference between  $i = 1$  and  $i = 2$  (which implies a statistical difference between  $i = 1$  and  $i > 1$ ).

found benefits in using long walks<sup>15</sup>. Overall, the assumption made by the cross-media method is better than the one underlying the random walk technique.

If we focus on Table 3, we can also compare the cross-media and the generalized diffusion processes. Both methods use the same number of nearest neighbors  $k^*$  but the former one makes only one iteration ( $i = 1$ ) while the latter one makes iterations until convergence. We can observe that in that case too, very short walks better perform than long walks.

However, by comparing the last rows of Table 3 and Table 2, we can conclude that the generalized diffusion process also outperforms the random walk method. Accordingly, it is better to take into account a small set of nearest neighbors in the diffusion process by using the operator  $\mathbf{K}$  with  $k \ll l$  instead of  $k = l$ . These experiments suggest that it is better to use short walks and small set of nearest neighbors in our multimedia retrieval model.

The results shown in Tables 2 and 3 also allow us to comment on the different search scenarios. Indeed, in this setting  $x_{(i)}^{qt}$  is like having a text query only and the graph based technique propagates textual relevance scores to visual similarities. This case is the one that has been considered so far in the experiments using random walk based techniques and in the research works [31, 30] in particular. The results given by  $y_{(i)}^{qt}$  are, on the contrary, the case where we are given in addition to the text query an image query. In that case we use the semantically filtered visual relevance scores as a prior and the purpose of Eq. 26 is to propagate the latter measures to text based similarities. We observe that  $y_{(i)}^{qt}$  are superior to  $x_{(i)}^{qt}$  in terms of MAP measures. Hence, these results confirm that, as for cross-media similarities, the random walk technique can give better results when the user can express her information need by a multimedia query instead of a text query solely.

### 7.1.2 Impact of the number of nearest neighbor $k$

In what follows, we study the results provided by different settings using different values of  $k$ . Moreover, we focus on the impact of using a prior or not in Eqs. 25 and 26. Indeed, as suggested in [31], it is important to add a prior in order to avoid the random walk process getting trapped in sub-local optimal solutions independent of the query. Hence, we consider the following set of parameters :

- $k \in \{10, 30, 50, 100, l\}$  (with or without nearest neighbor operator)
- $\gamma \in \{0, 0.3, \gamma^*\}$  (with or without prior, where  $\gamma^*$  is the parameter value among  $\{0.1, 0.2, \dots, 0.9\}$  that gave the best performance)
- $\beta = 0$  (no late fusion of similarity matrices)

<sup>15</sup>However, the tasks addressed in [21] are different since the graphs they deal with are sparse.

		IAPR		WIKI10		WIKI11		WEBQ
		$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_t^{qt}$
		27.6	26.3	24	26.3	18	27.8	57
$\gamma$	$k$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$x_{(i)}^{qt}$
0	10	35.5	19.3 <sup>†</sup>	24 <sup>†</sup>	23.5 <sup>†</sup>	19.9 <sup>†</sup>	22.5 <sup>†</sup>	66.1 <sup>†</sup>
0.3	10	36.5	25	28.1	<b>29.9</b>	22.9	<b>31</b>	66.2
$\gamma^*$	10	<b>36.6</b>	26.9*	<b>28.2</b>	<b>29.9</b>	23.1	<b>31</b>	67.4*
0	30	34.9	22.2 <sup>†</sup>	25.7 <sup>†</sup>	23.7 <sup>†</sup>	21.4 <sup>†</sup>	19.9 <sup>†</sup>	68.3 <sup>†</sup>
0.3	30	35.9	26.4	27.9	29.8	23.2	30.6	66
$\gamma^*$	30	35.9	<b>27.1</b> *	28	29.8	23.2	30.7	67.9*
0	50	33.6	22.3 <sup>†</sup>	25.7 <sup>†</sup>	23 <sup>†</sup>	21.3 <sup>†</sup>	16.4 <sup>†</sup>	68.7 <sup>†</sup>
0.3	50	35.1	26.6	28	29.8	<b>23.3</b>	30.2	66
$\gamma^*$	50	35.1	26.9	<b>28.2</b>	<b>29.9</b>	<b>23.3</b>	30.2	68.2*
0	100	31.3 <sup>†</sup>	21.7 <sup>†</sup>	25.2 <sup>†</sup>	21.4 <sup>†</sup>	18.9 <sup>†</sup>	12.7 <sup>†</sup>	69.1 <sup>†</sup>
0.3	100	33.3	26.2	27.4	29.5	22.1	29.6	66.1
$\gamma^*$	100	33.5	26.5	27.6	29.7	22.3	29.6	68.5*
0	$l$	28.7 <sup>†</sup>	20.8 <sup>†</sup>	18.9 <sup>†</sup>	15.7 <sup>†</sup>	12.6 <sup>†</sup>	6.9 <sup>†</sup>	69.3 <sup>†</sup>
0.3	$l$	33.4	26.6	25.9	28.3	19.6	27.8	66.1
$\gamma^*$	$l$	33.4	26.6	25.9	28.3	19.9	28.2	68.7*

Table 4: **Varying the number of nearest neighbors  $k$  and the weight of the prior.** Results are shown for  $i = 1$ ,  $k \in \{10, 30, 50, 100, l\}$  and  $\gamma \in \{0, 0.3, \gamma^*\}$ , where  $\gamma^*$  was the best  $\gamma$  found in the set of  $\{0.1, 0.2, \dots, 0.9\}$ . Adding a prior always leads to significantly better results, as also shown by the symbol <sup>†</sup> indicating a statistical difference between  $\gamma = 0.3$  (often best or close to best) and  $\gamma = 0$ . In contrast, there is rarely a statistical difference between  $\gamma = 0.3$  and  $\gamma^*$  indicated by the symbol \*. Finally, if there is a statistical difference between  $k = 10$  and other  $k$  values the results of  $k > 10$  are colored in magenta.

The results using these different settings for  $i = 1$  are given in Table 4 and with  $i = \infty$  in Table 5. Note that the last rows with  $k = l$  correspond to the classical random walk ( $rw^{qt}$ ) while the other rows, with  $k \ll l$ , lead to the generalized diffusion model ( $gd^{qt}$ ) that integrates the nearest neighbor operator.

Analyzing these results, and excluding the case of WEBQ, we observe the followings :

- Concerning the  $k$  value, best or near best results are obtained with  $k = 10$  for any value of  $\gamma$ . When the best results are achieved with  $k > 10$ , the latter parameter is below or equal to 50 and the gain is neither high nor statistically different as compared to  $k = 10$ . Accordingly, we conclude that for both the cross-media and the generalized diffusion model using  $k \approx 10$  could be considered as a default setting of our multimedia relevance model.
- When  $\gamma = 0$ , the  $cm^{qt}$  approach with  $i = 1$  (shown in Table 4), always yields to much higher performances than the  $gd^{qt}$  method with  $i = \infty$  (shown in Table 5). While this is not always the case when we have  $\gamma > 0$ , the setting  $\gamma^*$ ,  $k = 10$  and  $i = 1$  yields to results that are close (and statistically not different) to the best obtained values. In other words, and as already suggested beforehand, when using graph based techniques, one should favor the cross-media oriented diffusion process along with a relatively small number of nearest neighbors as proxies.
- When we consider the random walk, it generally yields to worse results than  $cm^{qt}$  and  $gd^{qt}$  for all three datasets whatever the value of  $\gamma$ .

Note, that in the case of WEBQ, some of these observations seems to be less true, where increasing both  $k$  and  $i$  improves the search results in terms of MAP. Best results are obtained with  $k = 50$ ,  $\gamma = 0.1$  and  $i = \infty$ , which again shows that using the operator  $\mathbf{K}(\cdot, k)$  (but this time with a higher  $k$ ) is a good idea. Concerning the number of iterations, we obtain a nice improvement over  $i = 1$  and close to  $i = \infty$  with a single extra step<sup>16</sup>.

<sup>16</sup>With  $\gamma^* = 0.1$ , we obtain that  $\text{MAP} = \{69.8, 69.9, 69.8, 69.8, 69.7\}$  respectively for  $k \in \{10, 30, 50, 100, l\}$

		IAPR		WIKI10		WIKI11		WEBQ
		$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_t^{qt}$
		27.6	26.3	24	26.3	18	27.8	57
$\gamma$	$k$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$x_{(i)}^{qt}$
0	10	31.4 <sup>†</sup>	16.6 <sup>†</sup>	22.2 <sup>†</sup>	20.7 <sup>†</sup>	19.1 <sup>†</sup>	14 <sup>†</sup>	66.4
0.3	10	35.1	24.2	28.4	29.2	<b>22.8</b>	<b>31.3</b>	66
$\gamma^*$	10	<b>35.4</b>	26.8*	<b>28.4</b>	29.2	<b>23.6</b>	<b>31.3</b>	66.9*
0	30	29 <sup>†</sup>	14.6 <sup>†</sup>	22.1 <sup>†</sup>	<b>18.3</b> <sup>†</sup>	18.6	<b>10.2</b> <sup>†</sup>	69.1
0.3	30	33.4	24.1	27.8	<b>29.4</b>	22.1	30.9	67.3
$\gamma^*$	30	34.3*	26.9*	27.8	<b>29.4</b>	22.4	30.9	69.7
0	50	29.3	15.9 <sup>†</sup>	19.9 <sup>†</sup>	<b>18.1</b> <sup>†</sup>	15.1 <sup>†</sup>	<b>8.7</b> <sup>†</sup>	69.5 <sup>†</sup>
0.3	50	<b>30.7</b>	24.6	26.2	<b>29.4</b>	20.3	29.4	67.7
$\gamma^*$	50	33.1	<b>26.7</b> *	27.3	<b>29.4</b>	21	30	<b>70.4</b> *
0	100	<b>24.9</b> <sup>†</sup>	15.5 <sup>†</sup>	<b>18.9</b> <sup>†</sup>	<b>16.6</b> <sup>†</sup>	<b>10.8</b> <sup>†</sup>	<b>7.2</b> <sup>†</sup>	69.3 <sup>†</sup>
0.3	100	<b>30.2</b>	25.1	<b>25.5</b>	28.4	<b>17.3</b>	<b>28.5</b>	67.8
$\gamma^*$	100	32.3	26.4	<b>26.3</b>	28.4	20.5	<b>28.9</b>	70.3*
0	$l$	<b>15.4</b> <sup>†</sup>	<b>8.4</b> <sup>†</sup>	<b>17</b> <sup>†</sup>	<b>13.5</b> <sup>†</sup>	<b>11.3</b> <sup>†</sup>	<b>5.1</b> <sup>†</sup>	<b>68.4</b> <sup>†</sup>
0.3	$l$	31.9	<b>26.4</b>	<b>25.3</b>	27.9	19.1	<b>27.6</b>	<b>66.6</b>
$\gamma^*$	$l$	32.1	26.6	<b>25.5</b>	27.9	<b>19.8</b>	<b>28.2</b>	<b>69.5</b> *

Table 5: **Varying the number of nearest neighbors  $k$  and the weight of the prior.** Results are shown for  $i = \infty$ ,  $k \in \{10, 30, 50, 100, l\}$  and  $\gamma \in \{0, 0.3, \gamma^*\}$ , where  $\gamma^*$  was the best  $\gamma$  found in the set of  $\{0.1, 0.2, \dots, 0.9\}$ . Adding a prior always leads to significantly better results, as also shown by the symbol <sup>†</sup> indicating a statistical difference between  $\gamma = 0.3$  (often best or close to best) and  $\gamma = 0$ . In contrast, there is rarely a statistical difference between  $\gamma = 0.3$  and  $\gamma^*$  indicated by the symbol \* (except WEBQ where we always found that the best  $\gamma$  was 0.1). Finally, if there is a statistical difference between  $k = 10$  and other  $k$  values the results of  $k > 10$  are colored in magenta.

This shows that even if a single iteration is not the best option, using only few iterations and small  $k$  values are sufficient. This confirms the tendencies we have observed so far.

### 7.1.3 Impact of the prior weight $\gamma$

Next, we can observe the following facts in regard to the integration of a prior in the graph based methods ( $\gamma = 0$  vs.  $\gamma > 0$ .) The case  $\gamma > 0$ ,  $k = l$ ,  $i = \infty$ , is typical to the initial random walk technique suggested in [31, 30]. In contrast, the case  $\gamma = 0$ ,  $k \leq l$ ,  $i = 1$ , is related to the cross-media technique. However, setting  $\gamma > 0$  in the latter case adds a prior in the cross-media oriented diffusion process and such a model has not been evaluated so far.

From the three last rows in Table 5, we can deduce that adding a prior to the random walk approach ( $i = \infty$ ) generally improves the search results. This outcome is aligned with the findings in [31, 30] which showed that the prior towards textual relevance scores allows better search results. In the case of a symmetric search scenario, we can also compute  $y_{(i)}^{qt}$  and in that case too, adding a semantically filtered visual prior improves the random walk based diffusion process performances. We thus extend the findings provided in [31, 30] to the symmetric search scenario.

Yet,  $y_{\infty}^{qt}$  scores do not surpass  $x_{\infty}^{qt}$  ones except for the IAPR task. Accordingly, using an image query with the random walk technique does not necessarily improve the search results. This observation might partly explain why research works using random walks in multimedia fusion have not studied further the symmetric search scenario. Note nevertheless that the random walk ( $y_{\infty}^{qt}$ ) always outperforms the image reranking ( $s_v^{qt}$ ) score.

If we compare the effect of adding a prior in the cross-media oriented diffusion process ( $i = 1$ ), by setting  $\gamma > 0$ , we observe that in all cases introducing priors was beneficial. Concerning the weight  $\gamma$ , we found that in most cases  $\gamma = 0.3$  was either optimal or yielded to values close to optimal, except in the case of the WEBQ dataset, where  $\gamma = 0.1$  led to much better results. Therefore, in what follows we will set  $\gamma$  to 0.3 for IAPR

			IAPR		WIKI10		WIKI11	
			$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$	$s_v^{qt}$	$s_t^{qt}$
			27.6	26.3	24	26.3	18	27.8
$\beta$	$i$	$k$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$	$y_{(i)}^{qt}$	$x_{(i)}^{qt}$
0	1	10	36.5	25	28.1	<b>29.9</b>	22.9	31
0	$\infty$	10	35.1	24.2	28.4	29.2	22.8	<b>31.3</b>
0	$\infty$	$l$	31.9*	26.4*	25.3*	27.9*	19.1*	27.6*
0.5	1	10	36.2	<b>27.3</b>	27.1	29.3	<b>20.5</b>	<b>29.7</b>
0.5	$\infty$	10	34	<b>27.2</b>	27.3	29.2	21	<b>29.5</b>
0.5	$\infty$	$l$	30.7*	26	24.6*	28*	<b>15.8*</b>	27.8*
$\beta^*$	1	10	<b>36.6</b>	<b>27.3</b>	28.5*	<b>29.9*</b>	23	30.9
$\beta^*$	$\infty$	10	34.9	<b>27.2</b>	<b>29.2</b>	29.3	<b>24</b>	31.2
$\beta^*$	$\infty$	$l$	31.7*	26.4	25.5	28	18.8*	27.8*
1	1	10	<b>27.7</b>	<b>27.1</b>	<b>24.5*</b>	<b>26.6</b>	<b>16.6</b>	<b>27.2</b>
1	$\infty$	10	<b>26.3</b>	25.9	<b>24.4</b>	<b>26.3</b>	<b>15.2<sup>†</sup></b>	27
1	$\infty$	$l$	<b>28.1*</b>	26	<b>22.7</b>	<b>26.7</b>	<b>11.5*</b>	27.1

Table 6: **Combining the similarity matrices.** We vary  $\beta$  and show the obtained results for  $x_{(i)}^{qt}$  and  $y_{(i)}^{qt}$  with the prior  $\gamma = 0.3$ . The symbol  $\dagger$  indicates a statistical difference between them. When there is a statistical difference between  $\beta = 0$  and  $\beta > 0$  (all the other parameters remaining the same), the values of the results given by  $\beta > 0$  are colored in magenta.

and WIKI datasets and to 0.1 for WEBQ, and denote these respective settings by  $\hat{\gamma}$ .

Similarly, adding the prior to the generalized diffusion model  $gd^{qt}$  was always beneficial. Besides, while without the prior ( $\gamma = 0$ ) the  $cm^{qt}$  models outperform the  $gd^{qt}$  ones, this is less true when we set  $\gamma > 0$ . Indeed, if we add a prior to both techniques, we observe that the generalized diffusion models give similar results than the cross-media models and in some cases (WIKI11), we can even note some better results.

At this point of our experimental report, we can make the following comments on the proposed multimedia retrieval model :

- (i) The results given in Tables 4 and 5 support the fact that graph based methods are effective techniques to combine visual and textual information in CBMIR as long as the parameters of Eq. 25 and Eq. 26 are set correctly since we can obtain much better results than the baselines given by  $s_t^{qt}$  and  $s_v^{qt}$  respectively;
- (ii) Adding a prior in any of the presented diffusion process improved the results as compared to the ones that do not use any prior.
- (iii) The diffusion processes using nearest neighbors as proxies ( $cm^{qt}$  and  $gd^{qt}$ ) outperform the random walk oriented diffusion process for both  $y_{(i)}^{qt}$  and  $x_{(i)}^{qt}$ . We also claim that  $k \approx 10$  and  $\gamma = 0.3$  can be considered as default parameters values for Eq. 25 and Eq. 26.

Finally, Table 4 allows us to further compare the results obtained in an asymmetric search scenario  $x_{(i)}^{qt}$  against the ones obtained in a symmetric search scenario  $y_{(i)}^{qt}$ . Interestingly, even if we use the best settings, the MAP measures for  $y_{(i)}^{qt}$  do not always outperform the ones for  $x_{(i)}^{qt}$  (actually it outperforms only for IAPR). However, these observations do not mean that in CBMIR, multimedia queries are not necessarily useful and adding an image query to a text query does not help the search results significantly. Indeed, retrieved lists provided by  $y_{(i)}^{qt}$  and  $x_{(i)}^{qt}$  should not be considered as in competition with each other. They should be viewed as complementary lists of pseudo-relevant items and in that perspective,  $s_v^{qt}$  and  $s_t^{qt}$  too have to be considered as complementary pseudo-relevant top lists of multimedia objects. In the section 7.2, we will thus investigate the combination of those different scores in order to analyze to what extent these top lists are mutually complementary.

### 7.1.4 Combining the similarity matrices

We now examine if fusing semantically filtered visual and textual similarity matrices by setting  $\beta \neq 0$  in Eq. 25 and Eq. 26 is beneficial or not. Combining monomedia similarity matrices was indeed suggested in [31, 30] in the random walk case. Consequently, we study the impact of  $\beta$  in the following experiments. As a result of the previous experiments, we selected the following parameter values :

- $i = 1, k = 10$  (cross-media oriented diffusion process)
- $i = \infty, k = 10$  (generalised diffusion model)
- $i = \infty, k = l$  (random walk oriented diffusion process)

For each case, we set  $\gamma = 0.3$  and we varied  $\beta \in [0, 1]$ , but we only show three particular cases :  $\beta = 0, \beta = 1$  and  $\beta^* \in [0.1, 0.9]$  which correspond to the values that gave the best results for each model. Table 6 presents the different MAP values we obtained. Note that we do not show the results for the WEBQ task because we did not compute<sup>17</sup> the visual similarity matrix  $S_v^{qt}$  in this case. From Table 6 we can make the following observations :

- Adding  $S_v^{qt}$  to  $S_t^{qt}$  ( $\beta > 0$ ) was clearly beneficial only in a few cases, more precisely for  $cm_{tv}^{qt}$  and  $gd_{tv}^{qt}$  in the case of IAPR and for  $cm_{vt}^q$  and  $gd_{vt}^q$  in the case of WIKI11. However, in the latter cases, an equal weighting is far from optimal or even worse and the parameter  $\beta$  has to be properly set. As a consequence, we can say that, in general,  $\beta = 0$  gives the best or near optimal solutions and adding the similarity matrix of the complementary modality in Eqs. 25 and 26 does not bring any further gain.
- As previously, both the cross-media and the generalized diffusion model results improve over the baselines  $s_v^{qt}$  and  $s_t^{qt}$  scores, and remain significantly better than the classical random walk process.

Note that the particular case  $\beta = 1$  corresponds, either to a semantically filtered “monomodal” pseudo-relevance feedback ( $i = 1$ ), or to a semantically filtered “monomodal” diffusion process ( $i = \infty$ ). Compared to the opposite case,  $\beta = 0$ , which is a transmodal semantically filtered pseudo-relevance feedback, they perform in general worse. This actually shows again the interest of exploiting both modalities using graph based techniques with a transmedia principle.

## 7.2 Combination of different relevance scores and similarities in a late fusion scheme

In the next set of experiments, we follow the research works described in [15, 2, 3, 16] where the authors propose to combine text based scores, semantically filtered visual scores and graph based scores as for the final relevance scores. In that perspective, we first study the following cases :

$$rsv_{cm}^{qt,tv}(q, \cdot) = \alpha s_t^{qt}(q, \cdot) + (1 - \alpha) x_{(1)}^{qt} \quad (27)$$

$$rsv_{cm}^{qt,vt}(q, \cdot) = \alpha s_v^{qt}(q, \cdot) + (1 - \alpha) y_{(1)}^{qt} \quad (28)$$

$$rsv_{gd}^{qt,tv}(q, \cdot) = \alpha s_t^{qt}(q, \cdot) + (1 - \alpha) x_{\infty}^{qt} \quad (29)$$

$$rsv_{gd}^{qt,vt}(q, \cdot) = \alpha s_v^{qt}(q, \cdot) + (1 - \alpha) y_{\infty}^{qt} \quad (30)$$

### 7.2.1 Effect of adding the initial semantically filtered scores

Before a comparison between  $rsv_{cm}^{qt,vt}$  and  $rsv_{rw}^{qt,tv}$ , let us first compare again the case of  $\gamma = 0$  and  $\hat{\gamma} > 0$ , but this time in the context of Eqs. 27 and 28. The results are shown in Table 7 where we varied  $\alpha$  between 0 and 1 (with a 0.1 step), but we only show the results for  $\alpha = 0.5$  and  $\alpha^*$  (which is the value that outperforms all other ones).

From this table we can draw the following conclusions :

- Adding  $s_t^{qt}$  to  $x_{(1)}^{qt}$  (Eq. 27) is always a winning strategy whether  $\gamma = 0$  or  $\hat{\gamma} > 0$ .

<sup>17</sup>As we shall see, setting  $\beta > 0$  does not necessarily bring any improvements of the results for IAPR, WIKI10 and WIKI11. Therefore, since the computation of visual similarities in the case of WEBQ is heavy, we did not compute it.



			IAPR		WIKI10		WIKI11		WEBQ
			$s_v^{q_t}$	$s_t^{q_t}$	$s_v^{q_t}$	$s_t^{q_t}$	$s_v^{q_t}$	$s_t^{q_t}$	$s_t^{q_t}$
			27.6	26.3	24	26.3	18.1	27.8	57
$k$	$\alpha$	$\gamma$	$rsv_{cm}^{q_t,vt}$	$rsv_{cm}^{q_t,tv}$	$rsv_{cm}^{q_t,vt}$	$rsv_{cm}^{q_t,tv}$	$rsv_{cm}^{q_t,vt}$	$rsv_{cm}^{q_t,tv}$	$rsv_{cm}^{q_t,tv}$
10	0	0	35.5	19.3 <sup>†</sup>	24 <sup>†</sup>	23.5 <sup>†</sup>	19.9 <sup>†</sup>	22.5 <sup>†</sup>	64.4 <sup>†</sup>
10	0.5	0	35.6	<b>27</b>	28	30	22.7	30.9	66.9
10	$\alpha^*$	0	36.5	<b>27</b>	28.2	30	23.1	30.9	66.9
10	0	$\hat{\gamma}$	36.5	25 <sup>†</sup>	<b>28.1<sup>†</sup></b>	<b>29.9<sup>†</sup></b>	<b>22.9</b>	<b>31<sup>†</sup></b>	<b>66.9<sup>†</sup></b>
10	0.5	$\hat{\gamma}$	<b>34.5</b>	<b>27</b>	<b>26.9</b>	<b>28.8</b>	<b>21.8</b>	30	<b>66.2</b>
10	$\alpha^*$	$\hat{\gamma}$	<b>36.6*</b>	<b>27</b>	28.2*	29.9*	23.1*	<b>31*</b>	66.9*
30	0	0	34.9	22.2 <sup>†</sup>	25.7 <sup>†</sup>	23.7 <sup>†</sup>	21.4	19.9 <sup>†</sup>	68.3
30	0.5	0	35.1	26.7	27.8	<b>30.1</b>	22.9	30.7	67.9
30	$\alpha^*$	0	35.8*	26.9	28.2*	<b>30.1</b>	23.3	30.7	68.6*
30	0	$\hat{\gamma}$	35.9 <sup>†</sup>	<b>26.4</b>	<b>27.9<sup>†</sup></b>	<b>29.8<sup>†</sup></b>	<b>23.2<sup>†</sup></b>	<b>30.6<sup>†</sup></b>	67.9 <sup>†</sup>
30	0.5	$\hat{\gamma}$	<b>33.8</b>	27	<b>26.5</b>	<b>28.6</b>	<b>21.7</b>	29.6	66.5
30	$\alpha^*$	$\hat{\gamma}$	35.9*	27	<b>27.9*</b>	29.8*	23.3*	30.6*	67.9*
50	0	0	33.6	22.3 <sup>†</sup>	25.7 <sup>†</sup>	23 <sup>†</sup>	21.3	16.4 <sup>†</sup>	68.7 <sup>†</sup>
50	0.5	0	34.5	26.6	27.7	30	23.1	30.3	68.3
50	$\alpha^*$	0	34.9	26.9	<b>28.3</b>	30	<b>23.5*</b>	30.3	69*
50	0	$\hat{\gamma}$	35.1 <sup>†</sup>	<b>26.6<sup>†</sup></b>	<b>28<sup>†</sup></b>	<b>29.8<sup>†</sup></b>	23.3 <sup>†</sup>	<b>30.2<sup>†</sup></b>	<b>68.2<sup>†</sup></b>
50	0.5	$\hat{\gamma}$	<b>33.3</b>	26.9	<b>26.4</b>	<b>28.5</b>	<b>21.8</b>	29.3	<b>66.6</b>
50	$\alpha^*$	$\hat{\gamma}$	35.1*	26.9	28*	29.8*	<b>23.4*</b>	30.2*	<b>68.2*</b>
$l$	0	0	28.7 <sup>†</sup>	20.8 <sup>†</sup>	18.9 <sup>†</sup>	15.7 <sup>†</sup>	12.6 <sup>†</sup>	6.9 <sup>†</sup>	69.3
$l$	0.5	0	33.3	25.8	25.9	28	19.6	26.8	68.9
$l$	$\alpha^*$	0	33.3	26.7	25.9	28.3	19.7*	28.2*	<b>69.6</b>
$l$	0	$\hat{\gamma}$	<b>33.4<sup>†</sup></b>	<b>26.6</b>	<b>25.9</b>	<b>28.3</b>	19.6	27.8	<b>68.7</b>
$l$	0.5	$\hat{\gamma}$	<b>31.2</b>	26.6	25.3	27.7	19.6	28.2	<b>66.9</b>
$l$	$\alpha^*$	$\hat{\gamma}$	33.4*	26.7	25.9	28.3	19.6	28.2	<b>68.7</b>

Table 7: **Effect of adding the initial semantically filtered scores  $s_v^{q_t}$  and  $s_t^{q_t}$  to the cross-media scores.** We show the results of  $rsv_{cm}^{q_t,vt}$  and  $rsv_{cm}^{q_t,tv}$  with  $\beta = 0$ . The symbol <sup>†</sup> indicates a statistical difference between  $\alpha = 0.5$  and  $\alpha = 0$ . and the symbol \* indicates a statistical difference between  $\alpha^*$  and  $\alpha = 0.5$ . Finally, when there is a statistical difference between corresponding  $\gamma = 0$  and  $\hat{\gamma} > 0$ , the results of  $\hat{\gamma}$  are colored in magenta.

- Adding  $s_v^{q_t}$  to  $y_{(1)}^{q_t}$  (Eq. 28) is not always beneficial and could be damaging for the search results especially with a non optimal  $\alpha$  value. However, we have to make the distinction between the cross-media oriented diffusion process with a prior ( $\hat{\gamma} > 0$ ) and the one without a prior ( $\gamma = 0$ ). In the former case, the graph based scores already benefited from  $s_v^{q_t}$  (used as a prior). As a result, adding the latter semantically filtered visual relevance score to  $y_{(1)}^{q_t}$  by means of a late fusion strategy does not bring any improvement or worse, it could hurt the performances which is typically the case for  $\alpha = 0.5$ . On the contrary, when we do not use any prior in the cross-media oriented diffusion process ( $k \leq 30$ ,  $\gamma = 0$ ), then we always observe a dramatic increase of the MAP measures.
- For both Eq. 27 and Eq. 28,  $\gamma = 0$  is in general better than  $\hat{\gamma} > 0$ , when  $\alpha = 0.5$  but compared to (the best)  $\alpha^*$ , the results are in general very similar (except for WEBQ) and with no statistical difference between them.
- It is difficult to judge between the two following cases :  $\gamma = 0$ ,  $\alpha > 0$  on the one hand and  $\hat{\gamma} > 0$  and  $\alpha = 0$  on the other hand. As regard to the cross media diffusion process, the former case promotes no prior but a late fusion while the latter case rather supports the integration of a prior in the diffusion process and no further linear combination. If we compare the latter case,  $\hat{\gamma} > 0$  and  $\alpha = 0$ , to the former one but with  $\gamma = 0$  and  $\alpha = 0.5$  (no need to tune the  $\alpha$  parameter), we see a slight advantage of the first setting over the second one. Therefore in what follows, we will pursue the experiments with the cross-

		IAPR		WIKI10		WIKI11		WEBQ
		$s_v^{q_t}$	$s_t^{q_t}$	$s_v^{q_t}$	$s_t^{q_t}$	$s_v^{q_t}$	$s_t^{q_t}$	$s_t^{q_t}$
		27.6	26.3	24	26.3	18.1	27.8	57
$k$	$\alpha$	$rsv_{gd}^{q_t, vt}$	$rsv_{gd}^{q_t, tv}$	$rsv_{gd}^{q_t, vt}$	$rsv_{gd}^{q_t, tv}$	$rsv_{gd}^{q_t, vt}$	$rsv_{gd}^{q_t, tv}$	$rsv_{gd}^{q_t, tv}$
10	0	35.1	24.2 <sup>†</sup>	28.4	29.2	22.8	31.3	67.4 <sup>†</sup>
10	0.5	35.1	<b>27</b>	27.5	28.7	22.9	30.4	<b>66.8</b>
10	$\alpha^*$	<b>36.1</b>	<b>27</b>	<b>28.7*</b>	29.2	<b>23.5</b>	<b>31.4*</b>	67.4*
30	0	33.4	<b>24.1<sup>†</sup></b>	27.8	<b>29.4</b>	22.1	30.9	<b>69.7<sup>†</sup></b>
30	0.5	33.5	26.1	27.6	28.8	23	30.2	<b>68.9</b>
30	$\alpha^*$	34.3	26.6	28.2	<b>29.4</b>	23.2	30.9	<b>69.7*</b>
50	0	<b>30.7</b>	<b>24.6<sup>†</sup></b>	<b>26.2</b>	<b>29.4</b>	20.3	29.4	<b>70.4<sup>†</sup></b>
50	0.5	31.6	26.2	26.7	28.8	21.6	29.4	<b>69.4</b>
50	$\alpha^*$	31.8	26.6	<b>26.8</b>	<b>29.4</b>	21.6	29.5	<b>70.4*</b>
$l$	0	<b>31.9</b>	26.4	25.3	<b>27.9</b>	19.1	27.6	<b>69.5<sup>†</sup></b>
$l$	0.5	<b>30.9</b>	26.5	<b>25.1</b>	<b>27.5</b>	19.6	<b>27.9</b>	<b>67.9</b>
$l$	$\alpha^*$	<b>32*</b>	26.6	<b>25.4</b>	<b>27.9</b>	19.8	28.1	<b>69.5*</b>

Table 8: **Effect of adding the initial scores  $s_v^{q_t}$  and  $s_t^{q_t}$  to  $gd^{q_t}$  and  $rw^{q_t}$ .** We show results of  $rsv_{gd}^{q_t, vt}$  and  $rsv_{gd}^{q_t, tv}$  with  $\beta = 0$ ,  $i = \infty$  and varying  $k$ . Note that  $k = l$  correspond to the  $rsv_{rw}^{q_t}$  case. The symbol <sup>†</sup> indicates a statistical difference between  $\alpha = 0.5$  and  $\alpha = 0$  and the symbol  $\star$  indicates a statistical difference between  $\alpha^*$  and  $\alpha = 0.5$ . Finally when there is a statistical difference between the corresponding cross-media oriented diffusion process ( $i = 1$  shown in Table 7) obtained with  $\hat{\gamma}$ , the results are colored in magenta.

media approach with a prior  $\hat{\gamma} > 0$  and no late fusion  $\alpha = 0$ . Nevertheless the other approach can also be considered as a possible multimedia retrieval model. Note that these observations support the fact that the cross-media approach gives search results that are complementary to the initial semantically filtered scores and this complementarity can be formulated *via* an early integration of these scores in the diffusion process as priors or *via* a late integration through a linear combination with the obtained graph based scores.

Accordingly, when using Eqs. 27 and 28 as multimedia retrieval models, we recommend to employ by default either the parameter setting  $\{k = 10, \gamma = \hat{\gamma}, \alpha = 0\}$  or  $\{k = 10, \gamma = 0, \alpha = 0.5\}$ . These configurations indeed lead to near optimal results on the first three datasets, and significantly better MAP values than the ones obtained with  $s_v^{q_t}$  and  $s_t^{q_t}$  for all datasets. Concerning, the WEBQ the best results we obtained were with  $k = l$ ,  $\gamma = 0$  and  $\alpha^*$ .

In Table 8, we show the performances of the other graph based techniques we are interested in : the generalized diffusion model and in particular, the random walk process (last three rows with  $k = l$ ). Hence, comparing Tables 7 and 8 allows us to assess Eqs. 27 and 28 against Eqs. 29 and 30. In other words, we again compare the two particular cases  $i = 1$  and  $i = \infty$  for this new set of experiments. Note that both  $rsv_{gd}^{q_t}$  and  $rsv_{rw}^{q_t}$  performed poorly without the prior so for these models, we did not make any experiment with  $\gamma = 0$  (instead we take  $\hat{\gamma} > 0$ ).

The comparison between the two cases,  $rsv_{gd}^{q_t}$  ( $k \ll l$ ) and  $rsv_{rw}^{q_t}$  ( $k = l$ ), yields to the following observations :

- As we already pointed out in Table 7, adding  $s_v^{q_t}$  and  $s_t^{q_t}$  as priors did not bring any improvement in most cases and with  $\alpha = 0.5$  we can even observe a decrease in the performances. Hence, when  $rsv_{gd}^{q_t}$  or  $rsv_{rw}^{q_t}$  are used, it is better to not recombine these scores again with the semantically filtered monomodal scores but to use directly  $gd^{q_t}$  and  $rw^{q_t}$  as multimedia retrieval models.
- When  $k = 10$  and  $\gamma = \hat{\gamma}$ ,  $rsv_{gd}^{q_t}$  ( $i = \infty$ ) and  $rsv_{cm}^{q_t}$  ( $i = 1$  shown in Table 7) yield very similar results (except for WEBQ), however for larger  $k$  values the  $rsv_{cm}^{q_t}$  outperforms  $rsv_{gd}^{q_t}$  showing again that for  $rsv_{gd}^{q_t}$  it is even more important to use small  $k$  values than for  $rsv_{cm}^{q_t}$ .
- For all datasets,  $rsv_{gd}^{q_t}$  outperforms  $rsv_{rw}^{q_t}$  (except WEBQ with  $k = 10$ ). More interestingly, the results obtained with  $rsv_{rw}^{q_t}$  remain significantly below the results provided by  $rsv_{cm}^{q_t}$  on the first three datasets.



Figure 9: Top retrieved images with textual similarity (second row), with cross-media  $rsv_{cm}^{q_t, tv}$  using  $k = 10$  and  $i = 1$  (third row) and with random walk  $rsv_{rw}^{q_t, tv}$  (last row), for the topic 7 at ImageCLEF Wikipedia Challenge 2010 (shown in first row). Green means relevant, red non-relevant.

Concerning WEBQ,  $rsv_{rw}^{q_t, tv}$  is better than  $rsv_{cm}^{q_t, tv}$  with  $k \leq 50$ , however for  $k = l$  the performances are comparable, especially when we consider  $\gamma = 0$  for  $rsv_{cm}^{q_t}$ . We thus state that cross-media oriented diffusion processes provide search results that are more complementary to  $s_t^{q_t}$  and  $s_v^{q_t}$  than random walk based diffusion processes since a late fusion (even with equal weighting) gives better results in the former case than in the latter case. This is further illustrated in Figure 9 on a Wikipedia query.

Finally, these two latter tables enable us to conclude about the asymmetric search scenario. Our experimental results show that, when given a text query only, the best multimedia fusion strategies are either to consider the cross-media oriented diffusion process (with or without prior) linearly combined with the initial semantically filtered text relevance scores which leads to the retrieval model  $rsv_{cm}^{q_t}$  (Eq. 27), or to use instead of  $rsv_{cm}^{q_t}$ , the generalized diffusion model  $gd^{q_t}$  without recombination with the initial scores. In both cases, it is important to use a relatively small  $k$  (e.g.  $k = 10$ ).

### 7.2.2 Combining all relevance scores in a late fusion scheme

Finally, we study the combination of both initial relevance scores  $s_v^{q_t}$  and  $s_t^{q_t}$  with both multimedia graph based scores  $y_{(i)}^{q_t}$  and  $x_{(i)}^{q_t}$  in an ultimate linear combination. This is possible in the case of the symmetric search scenario. We previously presented in section 5, Eq. 16 and Eq. 20 which refer to such combinations. For convenience, we recall these latter formulas below :

$$\begin{aligned} rsv_{cm}^{q_t}(q, \cdot) &= \alpha_t s_t^{q_t}(q, \cdot) + \alpha_v s_v^{q_t}(q, \cdot) + \alpha_{tv} cm_{tv}^{q_t}(q, \cdot) + \alpha_{vt} cm_{vt}^{q_t}(q, \cdot) \\ rsv_{gd}^{q_t}(q, \cdot) &= \alpha_t s_t^{q_t}(q, \cdot) + \alpha_v s_v^{q_t}(q, \cdot) + \alpha_{tv} gd_{tv}^{q_t}(q, \cdot) + \alpha_{vt} gd_{vt}^{q_t}(q, \cdot) \\ rsv_{rw}^{q_t}(q, \cdot) &= \alpha_t s_t^{q_t}(q, \cdot) + \alpha_v s_v^{q_t}(q, \cdot) + \alpha_{tv} rw_{tv}^{q_t}(q, \cdot) + \alpha_{vt} rw_{vt}^{q_t}(q, \cdot) \end{aligned}$$

where  $cm_{tv}^{q_t} = x_{(1)}^{q_t}$ ,  $cm_{vt}^{q_t} = y_{(1)}^{q_t}$ ,  $gd_{tv}^{q_t} = x_{\infty}^{q_t}$ ,  $gd_{vt}^{q_t} = y_{\infty}^{q_t}$  with  $k \ll l$  and  $rw_{tv}^{q_t} = x_{\infty}^{q_t}$ ,  $rw_{vt}^{q_t} = y_{\infty}^{q_t}$  with  $k = l$ .

According to the results we underlined previously, we chose the following settings :

- $k = 10$  (except for  $rw$  where  $k = l$ )
- $\beta = 0$  (no late fusion of similarity matrices).
- $\gamma = 0.3$  (using a monomodal prior) or  $\gamma = 0$  (without monomodal prior).
- $\alpha$  are set to uniform weights or to best performing weights.

We show the obtained results in Table 9. In addition, we show the MAP values obtained by the late fusion of semantically filtered relevance scores,  $\alpha_v s_v^{qt} + \alpha_t s_t^{qt}$ , which represents our baseline. From this table, we can draw the following conclusions :

- In a symmetric search setting, combining initial filtered scores with graph based diffusion processes scores is beneficial and performs better than the baseline. The graph based measures are thus complementary to the initial scores.
- $rsv_{cm}^{qt}$ ,  $rsv_{gd}^{qt}$  provide similar performances and outperform the random walk  $rsv_{rw}^{qt}$  method.

	IAPR	WIKI10	WIKI11
$0.5s_v + 0.5s_t$	34.5	35.2	35.4
$\alpha_v^* s_v + \alpha_t^* s_t$	35.4	35.2	35.4
$rsv_{cm}^{qt}, (\gamma = 0)$	37.3 <sup>†</sup>	35.9	33.6 <sup>†</sup>
$rsv_{cm}^{qt*}, (\gamma = 0)$	39.4	36.1	35.7
$rsv_{cm}^{qt}, (\gamma = 0.3)$	37.3 <sup>†</sup>	36.1	36
$rsv_{cm}^{qt*}, (\gamma = 0.3)$	39.5	36.1	36
$rsv_{gd}^{qt}, (\gamma = 0.3)$	37.3 <sup>†</sup>	35.8	35.1
$rsv_{gd}^{qt*}, (\gamma = 0.3)$	38.7	36	35.8
$rsv_{rw}^{qt}, (\gamma = 0.3)$	34.4	35.4	34.2 <sup>†</sup>
$rsv_{rw}^{qt*}, (\gamma = 0.3)$	36.1	35.6	35.4

Table 9: **Combining all relevance scores in a late fusion scheme.** We show the results of  $rsv^{qt}$  with uniform weights and  $rsv^{qt*}$  corresponding to the best linear combinations of the different scores. We considered  $k = 10$  (except for  $rsv_{rw}^{qt}$ ),  $\beta = 0$  and  $\gamma = 0.3$ . In addition as we average with the monomodal score, we also show the case of  $\gamma = 0$  (no prior) for the cross-media diffusion process. We do not show results with  $\gamma = 0$  for the random walk and generalized diffusion model as we have seen that iterating without the prior yields to much worse results. The symbol <sup>†</sup> indicates a statistical difference between uniform weights and tuned weights. We colored in magenta the values of the results of  $rsv^{qt}$  and the ones of  $rsv^{qt*}$  when there is a statistical difference with the semantically filtered late fusion (with uniform respectively tuned  $\alpha$  weights).

### 7.3 Advantage of the cross-media oriented diffusion process

With regard to the comparison between the cross-media and the random walk views of the proposed unified multimedia retrieval model, the experiments we conducted favors the former orientation as compared to the latter one. Indeed, all along this current section we have shown that :

- Choosing the probability distribution according to the semantically filtered relevance scores  $s_t^{qt}$  and  $s_v^{qt}$  as an initialization of the diffusion processes in Eqs. 25 and 26 respectively is better than the uniform distribution and this also supports the transmedia fusion principle in CBMIR.
- Using a small neighborhood as proxy in the transmedia approach by using  $\mathbf{K}(\cdot, k)$  with  $k \approx 10$  is in general better than choosing a large neighborhood; this is even more important when we iterate the generalized diffusion model;
- One-step or occasionally two-step walks provide better performances than longer walks when we do not use the initial scores as priors.
- The cross-media oriented diffusion process without the prior is the most complementary to the initial semantically filtered relevance scores and yields to better results than the random walk oriented diffusion process with or without the prior.
- Adding a prior in the diffusion processes is beneficial but using the latter information in a late fusion scheme after having calculated the graph based scores using only a single step gives comparable results.

Both solutions can be considered, however, the  $\gamma$  parameter for the former seems to be more stable than the best  $\alpha$  value in the latter case. Concerning the generalized diffusion model it is important to use the prior and preferably not to combine it with the semantically filtered monomodal scores.

- (vi) Overall, the best parameter setting we obtained as regard to our multimedia retrieval system is a late fusion between semantically filtered scores and cross-media oriented diffusion processes scores, the latter using the monomodal scores also as prior.

All these experiments allow us to claim that cross-media diffusion processes give better search results than random walk diffusion processes both for the asymmetric and symmetric search settings as long as the parameters are set correctly and in that perspective, we have also provided sets of parameters values one should use by default.

Below, we further underline and summarize two other important advantages of the cross-media framework compared to the random walk and the generalized diffusion model from the practical standpoint :

- **Complexity.** Our experimental results have shown that cross-media oriented diffusion processes give better results than random walk diffusion processes. As a consequence, we do not need to iterate the diffusion process until convergence and we have shown that a single step was sufficient to reach good performances. Hence, the model underlying the cross-media similarities not only provide better results but it also reduces the computation time. Moreover, when we use the cross-media settings ( $\beta = 0, k = 10, i = 1$ ), we employ the nearest neighbor thresholding operator  $\mathbf{K}(\cdot, k)$  on the initial monomodal scores. In that case, we do not even need to compute the  $l \times l$  similarity matrices, but only  $k \times l$  similarity values which correspond to the rows of  $S_t^{qt}$  and  $S_v^{qt}$  associated to the  $k$  nearest neighbors (even when using the prior these values are to be added simply to the corresponding rows). This further reduces the computational cost of the proposed graph based method. Since image representation and similarities have much higher time and storage complexities than text, overall, our proposal can easily tackle large collections of multimedia objects.
- **Score Normalization** In order to compare the two methods we have been interested in, we normalized the scores and similarities in order to have probability distributions. This normalization was necessary for the random walk approach as explained in section 5. However, since we have shown that iterating the diffusion process more than once was generally damaging, we concluded that we needed to iterate Eq. 25 and Eq. 26 only one time. As a result, it is no more mandatory to have probability distributions as for  $x_{(i)}^{qt}$  and  $y_{(i)}^{qt}$ . We thus computed  $rsv_{cm}^{qt}$  using another normalization : we replaced each score or similarity  $s^{qt}(d, d')$  by  $(s^{qt}(d, d') - \min\{s^{qt}(d, \cdot)\}) / (\max\{s^{qt}(d, \cdot)\} - \min\{s^{qt}(d, \cdot)\})$ . We show in Table 10<sup>18</sup>, the performances of the cross-media oriented diffusion process using this other normalization procedure. This is denoted by  $rsv_{cm}^{qt, nm}$ . Overall, we reached even better MAP results. As a consequence, this suggests that our study of graph based methods could be further enriched with the impact of other kinds of normalization methods.

	IAPR	WIKI10	WIKI11	WEBQ
$\alpha_v^* s_v + \alpha_t^* s_t$	35.4	35.2	35.4	57
$rsv_{rw}^{qt*}$	36.1	35.6	35.4	69.5
$rsv_{gd}^{qt*}$	38.7	36	35.8	70.4
$rsv_{cm}^{qt*}$	39.5	36.1	<b>36</b>	69.6
$rsv_{cm}^{qt, nm}$	<b>40.2</b>	<b>36.2</b>	35.6	<b>70.7</b>

Table 10: Results with different score normalization.

## 8 Conclusion

We have addressed the problem of multimedia information fusion in CBMIR and compared the cross-media similarities to the random walk methods. First of all, we have proposed a unifying framework that integrates the

<sup>18</sup>Note that for WEBQ, we do not have the image query and we can not provide the final scores  $rsv_{rw}^{qt}$  and  $rsv_{cm}^{qt}$  given by Eq. 20 and Eq. 16 unlike for other tasks. In that case, we show the best values we obtained among all the previous experimental results we presented. Note that we show single step ( $i = 1$ ) results in the case of  $rsv_{cm}^{qt, nm}$ .

text query based semantic filtering of multimedia scores and similarities and which generalizes both graph based techniques in a unifying model.

Furthermore, we have extensively studied many factors that impact the performances of these graph based methods. One of our goals was to provide some guidelines on how to best use those methods for two different multimodal search scenarios : the asymmetric and symmetric cases. Our findings have been validated on three real-world datasets which are public and accessible to the research community.

All in all, we can summarize our findings about graph based methods as follows :

- The text query based semantic filtering is an efficient fusion method that allows one to restrain the search space to multimedia items that are the most semantically related to the text query. We suggest to apply this first level of fusion before moving forward with graph based methods.
- Cross-media similarities and random walk based approaches can be seamlessly embedded into a unifying framework. The latter general graph based method is defined by Eq. 25 and Eq. 26 which allow one to take into account both the asymmetric and the symmetric search scenarios. The unifying framework exhibits transmedia diffusion processes with or without priors and bring to light the main differences between the two types of methods used by the community. But in a more general scope, it allows us to formalize the interesting features and parameters one should pay attention to when using an unsupervised graph based approach in content based image/text multimedia retrieval tasks.
- The experiments we conducted globally show that cross-media oriented diffusion processes outperform random walk based methods. Typically, we claim that the default setting for Eq. 25 and Eq. 26 which yields to near best performances on average are :
  - $\beta = 0$  (no late fusion between similarity matrices).
  - $\gamma = 0.3$  (using a prior helps)
  - $k \approx 10$  (a few nearest neighbors as proxies is better)
  - $i = 1$  (one iteration is sufficient)

We have shown that the cross-media oriented diffusion process being complementary to the initial relevance scores  $s_v^{qt}(q, \cdot)$  respectively  $s_t^{qt}(q, \cdot)$  can be successfully combined with them to further improve the retrieval accuracy. Overall, we obtained the best search results with  $rsv_{cm}^{qt}(q, \cdot) = \alpha_t s_t^{qt}(q, \cdot) + \alpha_v s_v^{qt}(q, \cdot) + \alpha_{tv} cm_{tv}^{qt}(q, \cdot) + \alpha_{vt} cm_{vt}^{qt}(q, \cdot)$  for all real-world tasks we tested especially when we used another normalization than the ones required by the iterative processes. Last but not least, cross-media oriented diffusion processes have a lower computational cost compared to both the random walk oriented and generalized diffusion models and hence such graph based techniques can tackle large multimedia repositories in a scalable way.

## A Text representation and similarities

Standard pre-processing techniques were first applied to the textual part of the documents. After stop-word removal, words were lemmatized and the collection of documents indexed with Lemur<sup>19</sup>.

We describe here the Lexical Entailment (LE) model used on the Wikipedia dataset as it is a less well-known model. [7] addressed the problem of IR as a statistical translation problem with the well-known noisy channel model. This model can be viewed as a probabilistic version of the generalized vector space model. The analogy with the noisy channel is the following one : to generate a query word, a word is first generated from a document and this word then gets “corrupted” into a query word. The key mechanism of this model is the probability  $p(v|u)$  that term  $u$  is “translated” by term  $v$ . These probabilities enable us to address a vocabulary mismatch, and also some kinds of semantic enrichments.

Then, the problem lies in the estimation of such probability models. We refer here to a previous work [20] on LE models to estimate the probability that one term entails another. It can be understood as a probabilistic term similarity or as a unigram LM associated to a word (rather than to a document or a query). Let  $u$  be a term in the corpus, then LE models compute a probability distribution over terms  $v$  of the corpus denoted by  $p(v|u)$ .

<sup>19</sup><http://www.lemurproject.org/>

These probabilities can be used in IR models to enrich queries and/or documents and to give a similar effect to the use of a semantic thesaurus. However, LE is purely automatic, as statistical relationships are only extracted once from the considered corpus. In practice, a sparse representation of  $p(v|u)$  is adopted, where we restrict  $v$  to be one of the 10 terms that are the closest to  $u$  using an information gain metric.

More formally, an entailment or similarity between words, expressed by a conditional probability  $p(v|u)$ , can be used to rank documents according to the following formula :

$$s_t(d_t, d'_t) = p(d_t|d'_t) = \prod_{v \in d_t} \sum_u p(v|u)p(u|d'_t). \quad (31)$$

where  $d_t$  (or  $q_t$ ) and  $d'_t$  are two texts,  $p(u|v)$  may be obtained by any of the methods described in [20] and  $p(u|d'_t)$  is the LM of  $d'_t$ .

Note that this model was essentially rediscovered in [33] and give substantial improvements compared to standard retrieval models (language models, divergence from randomness, information models). For instance, the LE model obtains a MAP of 26.3% compared to 22.6% on the 2010 Wikipedia dataset. Similarly, on the 2011 dataset, the LE MAP is 27.82% compared to a 24.3% an information based model[19].

## B Image representation and similarities

As for image representations, we used the Fisher Vector (FV), proposed in [48], an extension of the popular Bag-of-Visual word (BOV) image representation [56, 22], where an image is described by a histogram of quantized local features. The Fisher Vector, similarly to the BOV, is based on an intermediate representation, the visual vocabulary, which is built on the top of the low-level feature space. In our experiments we used two types of low-level features, the SIFT-like Orientation Histograms (ORH) and the local color (RGB) statistics (LCS) proposed in [18] and built an independent visual vocabulary for both of them.

The visual vocabulary was modeled by a Gaussian Mixture model (GMM)  $p(u|\lambda) = \sum_{i=1}^N w_i \mathcal{N}(u|\mu_i, \Sigma_i)$ , where  $\lambda = \{w_i, \mu_i, \Sigma_i; i = 1, \dots, N\}$  is the set of all parameters of the GMM and each Gaussian corresponds to a visual word. In the case of BOV representation, the low-level descriptors  $\{u_t; t = 1, \dots, T\}$  of an image  $d_v$ , are transformed into a high-level  $N$  dimensional descriptor,  $\gamma(d_v)$ , by accumulating over all low-level descriptors and for each Gaussian, the probabilities of generating a descriptor :

$$\gamma(d_v) = [\sum_{t=1}^T \gamma_1(u_t), \sum_{t=1}^T \gamma_2(u_t), \dots, \sum_{t=1}^T \gamma_N(u_t)] \quad (32)$$

where

$$\gamma_i(u_t) = \frac{w_i \mathcal{N}(u_t|\mu_i, \Sigma_i)}{\sum_{j=1}^N w_j \mathcal{N}(u_t|\mu_j, \Sigma_j)}. \quad (33)$$

The Fisher Vector [48] extends this BOV representation by going beyond counting measures (0-order statistics) and by encoding statistics (up to the second order) about the distribution of local descriptors assigned to each visual word. It rather characterizes the low-level features  $\{u_t\}_{t=1, \dots, T}$  of an image  $d_v$  by its deviation from the GMM distribution :

$$G_\lambda(d_v) = \frac{1}{T} \sum_{t=1}^T \nabla_\lambda \log \left\{ \sum_{j=1}^N w_j \mathcal{N}(u_t|\mu_j, \Sigma_j) \right\} \quad (34)$$

To compare two images  $d_v$  (or  $q_v$ ) and  $d'_v$  from two multimedia documents  $d$  (or respectively the query  $q$ ) and  $d'$ , a natural kernel on these gradients is the Fisher Kernel [48] :

$$s_v(d_v, d'_v) = G_\lambda(d_v)^\top F_\lambda^{-1} G_\lambda(d'_v), \quad (35)$$

where  $F_\lambda$  is the Fisher Information Matrix. As  $F_\lambda^{-1}$  is symmetric and positive definite, it has a Cholesky decomposition denoted by  $L_\lambda^\top L_\lambda$ . Therefore  $s_v(d_v, d'_v)$  can be rewritten as a dot-product between normalized vectors using the mapping  $\Gamma_\lambda$  with :

$$\Gamma_\lambda(d_v) = L_\lambda \cdot G_\lambda(d_v) \quad (36)$$

which we refer to as the Fisher Vector (FV) of the image  $d_v$ .

As suggested in [49], we further used a square-rooted and  $L2$  normalized versions of the FV and also built a spatial pyramid [38]. Regarding this latter point, we repeatedly subdivide the image into 1, 3 and 4 regions : we consider the FV of the whole image (1x1); the concatenation of 3 FV extracted for the top, middle and bottom regions (1x3) and finally, the concatenation of four FV one for each quadrants (2x2). In other words, the spatial pyramid (SP) we obtained for each image considering both LCS and ORH features is given by  $8 + 8 = 16$  FV. We used the dot product (linear kernel) to compute the similarity between the concatenation<sup>20</sup> of all FV for ORH and LCS.

## References

- [1] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J.M. Renders. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications*, 42(1):31–56, 2009.
- [2] J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J.M. Renders. XRCE’s participation to ImageCLEF 2008. In *Working Notes of CLEF 2008*, September 2008.
- [3] J. Ah-Pine, S. Clinchant, G. Csurka, and Yan Liu. XRCE’s participation to ImageCLEF 2009. In *Working Notes of the 2009 CLEF Workshop*, September 2009.
- [4] J. Ah-Pine, S. Clinchant, G. Csurka, F. Perronnin, and J-M. Renders. *Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval*, chapter 3.4. Volume INRE of etal [45], 2010.
- [5] Julien Ah-Pine, Stephane Clinchant, and Gabriela Csurka. Comparison of several combinations of multimodal and diversity seeking methods for multimedia retrieval. In *Multilingual Information Access Evaluation*, Lecture Notes in Computer Science (LNCS). Springer, 2010.
- [6] Joan Benavent, Xaro Benavent, Esther de Ves, Ruben Granados, and Ana Garcia-Serrano. Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches. In Brashler et al. [8].
- [7] Adam L. Berger and John D. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229. ACM, 1999.
- [8] Martin Brashler, Donna Harman, and Emanuele Pianta, editors. *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, 2010.
- [9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.
- [10] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998.
- [11] Eric Bruno, Nicolas Moënné-Loccoz, and Stéphane Marchand-Maillet. Design of multimodal dissimilarity spaces for retrieval of video documents. *PAMI*, 30(9):1520–1533, 2008.
- [12] Juan C. Caicedo, Jose G. Moreno, Edwin A. Niño, and Fabio A. González. Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Multimedia Information Retrieval*, 2010.
- [13] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.
- [14] S. Clinchant, J.-M. Renders, and G. Csurka. Trans-media pseudo-relevance feedback methods in multimedia retrieval. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, pages 569–576. Springer, 2008.

---

<sup>20</sup>Note that we do not need to explicitly concatenate all these vectors as  $\langle [u, v], [u', v'] \rangle = \langle u, u' \rangle + \langle v, v' \rangle$ .



- [15] S. Clinchant, J.M. Renders, and G. Csurka. XRCE's participation to ImageCLEF . In *CLEF Working Notes*, 2007.
- [16] Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.
- [17] Stéphane Clinchant, Gabriela Csurka, Julien Ah-Pine, Guillaume Jacquet, Florent Perronnin, Jorge Sánchez, and Keyvan Minoukadeh. Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
- [18] Stephane Clinchant, Gabriela Csurka, Florent Perronnin, and Jean-Michel Renders. XRCE's participation to ImageEval. In *ImageEval Workshop at CVIR*, 2007.
- [19] Stéphane Clinchant and Eric Gaussier. Information-based models for ad hoc IR. In *SIGIR*. ACM, 2010.
- [20] Stéphane Clinchant, Cyril Goutte, and Éric Gaussier. Lexical entailment for information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12*, pages 217–228, 2006.
- [21] Nick Craswell and Martin Szummer. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM.
- [22] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.
- [23] Gabriela Csurka and Stéphane Clinchant. An empirical study of fusion operators for multimodal image retrieval. In *CBMI*, 2012.
- [24] Gabriela Csurka, Stéphane Clinchant, and Adrian Popescu. Xrce's participation at wikipedia retrieval of imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.
- [25] Hugo Jair Escalante, Carlos A. Hernández, Luis Enrique Sucar, and Manuel Montes y Gómez. Late fusion of heterogeneous methods for multimedia image retrieval. In *MIR*, 2008.
- [26] Massimo Franceschet. Pagerank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6), 2011.
- [27] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on*, 22(1):363–376, 2013.
- [28] Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.
- [29] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia*, pages 35–44, 2006.
- [30] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. Reranking methods for visual search. *IEEE MultiMedia*, 14(3):14–22, 2007.
- [31] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. Video search reranking through random walk over document-level context graph. In *ACM Multimedia*, pages 971–980, 2007.
- [32] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 119–126, New York, NY, USA, 2003. ACM.

- [33] Maryam Karimzadehgan and ChengXiang Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *SIGIR*, pages 323–330. ACM, 2010.
- [34] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999.
- [35] Josip Krapac, Moray Allan, Jakob Verbeek, and Frédéric Jurie. Improving web-image search results using query-relative classifiers. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10)*, pages 1094–1101, San Francisco, United States, 2010. IEEE Computer Society.
- [36] Amy N. Langville and Carl D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM Rev.*, 47(1):135–161, January 2005.
- [37] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.
- [38] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [39] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009.
- [40] Hao Ma, Jianke Zhu, Michael R. Lyu, and Irwin King. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12(5):462–473, 2010.
- [41] João Magalhães and Stefan M. Rüger. An information-theoretic framework for semantic-multimedia retrieval. *ACM Trans. Inf. Syst.*, 28(4):19, 2010.
- [42] Nicolas Mailliot, Jean-Pierre Chevallet, and Joo-Hwee Lim. Inter-media pseudo-relevance feedback application to imageclef 2006 photo retrieval. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 735–738. Springer, 2006.
- [43] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. 1999.
- [44] Nobuyuki Morioka and Jingdong Wang. Robust visual reranking via sparsity and ranking constraints. In *ACM Multimedia*, pages 533–542, 2011.
- [45] H. Müller, P. Clough, Th. Deselaers, and B. Caputo, editors. *ImageCLEF- Experimental Evaluation in Visual Information Retrieval*, volume INRE. Springer, 2010.
- [46] Apostol (Paul) Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 991–1000, New York, NY, USA, 2007. ACM.
- [47] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Automatic multimedia cross-modal correlation discovery. In *KDD*, pages 653–658, 2004.
- [48] F. Perronnin and C. Dance. Fisher Kernels on visual vocabularies for image categorization. In *CVPR*. IEEE, 2007.
- [49] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the Fisher Kernel for large-scale image classification. In *ECCV*, 2010.
- [50] A. Popescu, T. Tsikrika, and J. Kludas. Overview Of The Wikipedia Retrieval Task At ImageCLEF 2010. In *Working notes of the 11th Workshop of the Cross-Language Evaluation Forum*. CLEF-campaign, September 2010.
- [51] Adrian Popescu. Télécom bretagne at ImageCLEF WikipediaMM 2010. In Braschler et al. [8].

- [52] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, 2010.
- [53] Sergio Rodriguez-Vaamonde, Lorenzo Torresani, and Andrew Fitzgibbon. What can pictures tell us about web pages?: improving document search using images. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 849–852, New York, NY, USA, 2013. ACM.
- [54] Stefan Rueger. *Multimedia Information Retrieval*. Morgan and Claypool Publishers, 1st edition, 2010.
- [55] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.
- [56] J. S. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.
- [57] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVID. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
- [58] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia*, pages 399–402, 2005.
- [59] Xinmei Tian, Linjun Yang, Jingdong Wang, Yichen Yang, Xiuqing Wu, and Xian-Sheng Hua. Bayesian video search reranking. In *ACM Multimedia*, pages 131–140, 2008.
- [60] A. Vinokourov, D. R. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. 2003.
- [61] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, G.-J. Qi, and Yan Song. Unified video annotation via multigraph learning. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(5):733–746, 2009.
- [62] Meng Wang, Xian-Sheng Hua, Jinhui Tang, and Richang Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *Multimedia, IEEE Transactions on*, 11(3):465–476, 2009.
- [63] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. Multimodal graph-based reranking for web image search. *Image Processing, IEEE Transactions on*, 21(11):4649–4661, 2012.
- [64] Xin-Jing Wang, Wei-Ying Ma, Gui-Rong Xue, and Xing Li. Multi-model similarity propagation and its application for web image retrieval. In *ACM Multimedia*, pages 944–951, 2004.
- [65] Peter Wilkins, Alan F. Smeaton, and Paul Ferguson. Properties of optimally weighted data fusion in cbmir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 643–650, New York, NY, USA, 2010. ACM.
- [66] Linjun Yang and Alan Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, pages 183–192, 2010.
- [67] Zheng-Jun Zha, Meng Wang, Jialie Shen, and Tat-Seng Chua. Text mining in multimedia. In *Mining Text Data*, pages 361–384. 2012.